Advances in Experimental Philosophy of Mind

Advances in Experimental Philosophy

Empirical and experimental philosophy is generating tremendous excitement, producing unexpected results that are challenging traditional philosophical methods. *Advances in Experimental Philosophy* responds to this trend, bringing together some of the most exciting voices in the field to understand the approach and measure its impact in contemporary philosophy. The result is a series that captures past and present developments and anticipates future research directions.

To provide in-depth examinations, each volume links experimental philosophy to a key philosophical area. They provide historical overviews alongside case studies, reviews of current problems and discussions of new directions. For upper-level undergraduates, postgraduates and professionals actively pursuing research in experimental philosophy these are essential resources.

New titles in the series include:
Advances in Experimental Epistemology, edited by James R. Beebe
Advances in Experimental Moral Psychology, edited by Hagop Sarkissian and Jennifer Cole Wright

Advances in Experimental Philosophy of Mind

Edited by
Justin Sytsma

Series: Advances in Experimental Philosophy

BLOOMSBURY
LONDON • NEW DELHI • NEW YORK • SYDNEY

Bloomsbury Academic
An imprint of Bloomsbury Publishing Plc

50 Bedford Square
London
WC1B 3DP
UK

1385 Broadway
New York
NY 10018
USA

www.bloomsbury.com

Bloomsbury is a registered trade mark of Bloomsbury Publishing Plc

First published 2014

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN: HB: 978-1-4725-1480-6
ePDF: 978-1-4725-0733-4
ePub: 978-1-4725-0772-3

Library of Congress Cataloging-in-Publication Data
Advances in experimental philosophy of mind/edited by Justin Sytsma.
pages cm. – (Advances in experimental philosophy)
Includes bibliographical references and index.
ISBN 978-1-4725-1480-6 (hardback) – ISBN 978-1-4725-0772-3 (epub) –
ISBN 978-1-4725-0733-4 (epdf) 1. Philosophy of mind–Research.
I. Sytsma, Justin, editor of compilation.
BD418.3.A325 2014
128'.2–dc23
2013041414

Typeset by Deanta Global Publishing Services, Chennai, India
Printed and bound in Great Britain

Notes on Contributors

Wesley Buckwalter is a postdoctoral researcher in the department of philosophy at the University of Waterloo in Ontario, Canada. Before that he completed his PhD in philosophy at the City University of New York, Graduate Center. His current research lies at the intersection of epistemology, psychology, and philosophy of cognitive science. He has authored or coauthored papers in *Noûs*, *Philosophical Studies*, *Mind & Language*, *Philosophical Psychology*, *Philosophical Topics*, *Philosophy Compass*, and the *Annual Review of Psychology* on a wide range of philosophical categories and concepts, including knowledge, belief, action, luck, intuitions, philosophical methodology, functionalism, phenomenal consciousness, emotion, and fiction.

Jared Friedman is an undergraduate at Case Western Reserve University studying cognitive science and philosophy (2014). Jared has worked in the Brain, Mind, and Consciousness Lab, under the guidance of Dr Anthony Jack, since he enrolled at Case (2012). Broadly stated, Jared's research interests lie at the intersection of experimental psychology and philosophy (particularly philosophy of mind and science). By coalescing philosophical insights with empirically informed theories of cognition, he aims to contribute to the understanding of folk-psychological phenomena, the cognitive origins of philosophical problems, and their associated psychological and neural mechanisms.

Chris Howard is a doctoral candidate in the department of philosophy at the University of Arizona. His current research focuses on issues in ethics (especially metaethics and practical reason) and epistemology.

Anthony (Tony) Jack leads the Brain, Mind and Consciousness Lab at Case Western Reserve University, and holds appointments in the departments of cognitive science (primary), philosophy, psychology, neurology, and neuroscience. After an influential first year studying physics and philosophy of science, he took his BA in psychology and philosophy from Oxford University,

and his PhD in experimental psychology from University College London. He then trained and worked in cognitive neuroscience at the Institute of Cognitive Neuroscience in Queen Square London, and then at the department of neurology, Washington University in St Louis. His work is directed at establishing an integrated science of the mind which better approximates a mature science, by being more theory driven, specific, and falsifiable. His work explores novel forms of interdisciplinary contact, using philosophical analysis to generate testable hypotheses, and cognitive neuroscience to constrain psychological theory.

Victor Kumar is currently a postdoctoral research fellow at the University of Michigan, Ann Arbor. He received his PhD in philosophy from the University of Arizona and works mainly at the intersection of ethics and empirical moral psychology.

Trevor Kvaran is director of information services at Communicus, a private market research firm. In this role, Trevor directs a team that is responsible for implementing and analyzing approximately 100 studies of advertising effectiveness each year for several Fortune 500 companies. Prior to starting with Communicus, Trevor completed a MA in philosophy through Georgia State University's brain and behavior program and a MA and PhD in psychology through the cognitive and neural systems program at the University of Arizona. Outside of his work with Communicus, Trevor's academic interests are in the psychology of charitable giving, moral psychology, and statistical computing.

Edouard Machery is professor in the department of history and philosophy of science at the University of Pittsburgh, a fellow of the Center for Philosophy of Science at the University of Pittsburgh, and a member of the Center for the Neural Basis of Cognition (University of Pittsburgh-Carnegie Mellon University). His research focuses on the philosophical issues raised by psychology and cognitive neuroscience with a special interest in concepts, moral psychology, the relevance of evolutionary biology for understanding cognition, the nature, origins, and ethical significance of prejudiced cognition, the foundation of statistics, and the methods of psychology and cognitive neuroscience. He is the author of *Doing without Concepts* (OUP, 2009) as well as the editor of *The Oxford Handbook of Compositionality* (OUP, 2012),

La Philosophie Expérimentale (Vuibert, 2012), *Arguing about Human Nature* (Routledge, 2013), and *Current Controversies in Experimental Philosophy* (Routledge, 2014).

Chris Meyers is currently associate professor of philosophy at the University of Southern Mississippi. He specializes in ethical theory, metaethics, moral psychology, and applied ethics. He is especially interested in applying empirical findings in the cognitive and behavior sciences to philosophical questions. Some of his recent published work includes "Defending moral realism from empirical evidence of disagreement," *Social Theory and Practice* 39 (2013) and "Brains, trolleys, and intuitions: defending deontology from the Greene/Singer argument," *Philosophical Psychology*, forthcoming.

Jennifer Nado is assistant professor of philosophy at Lingnan University, Hong Kong. She received her PhD from Rutgers University in 2011. Her primary field of research is metaphilosophy, focusing on the role of intuition in philosophical theorizing. Much of her work explores the epistemological impact of findings in experimental philosophy. She has published papers in *Philosophy and Phenomenological Research* and *The British Journal for the Philosophy of Science*.

Shaun Nichols is professor of philosophy at the University of Arizona. His books include *Sentimental Rules* (2004) and, coauthored with Stephen Stich, *Mindreading* (2003). His current research focuses on the psychological underpinnings of philosophical problems.

Mark Phelan is assistant professor of philosophy and cognitive science at Lawrence University in Wisconsin. Before coming to Lawrence, he was a postdoctoral fellow at Yale University. He holds a PhD in philosophy from the University of North Carolina, Chapel Hill. He specializes in the philosophies of language, mind, and psychology. Papers he has written or coauthored have appeared in journals such as *Mind & Language*, *Pacific Philosophical Quarterly*, *The British Journal of Aesthetics*, *Philosophical Studies*, and *Philosophical Topics*.

Kevin Reuter is a postdoctoral fellow at the philosophy department of the Ruhr University in Bochum. His research aims to understand the conceptual requirements that underwrite introspective processes. Another of his main

research projects is the theoretical and empirical investigation of the nature of pain and the semantics of pain reports. Publications include "Distinguishing the appearance from the reality of pain" (*Journal of Consciousness Studies*, 2011), "Is imagination introspective?" (*Philosophia*, 2011), "The importance of intentions in introspection" (forthcoming in *Austin on Language*).

Dustin Phillips is a graduate student in religious studies at Naropa University, studying Indo-Tibetan Buddhism.

Philip Robbins is an associate professor of philosophy at the University of Missouri in Columbia, where he directs the MU Experimental Philosophy Lab. He received his AB from Harvard University and his PhD from the University of Chicago. His primary research area is social cognition, with a special focus on mind perception and the psychology of moral categorization and judgment. He has published articles on a wide range of topics in the philosophy of mind and language, including mindreading, modularity, self-consciousness, and compositionality. He is also coeditor, with Murat Aydede, of the *Cambridge Handbook of Situated Cognition* (Cambridge University Press, 2009).

Justin Sytsma is a lecturer in the philosophy program at Victoria University of Wellington. His research focuses on issues in philosophy of psychology and philosophy of mind. As a practitioner of experimental philosophy, his research into these areas often involves the use of empirical methods. He is coauthor of a forthcoming volume on *The New Experimental Philosophy* from Broadview Press, in addition to authoring or coauthoring articles appearing in journals such as *Australasian Journal of Philosophy*, *Philosophical Studies*, *Philosophy of Science*, *Journal of Consciousness Studies*, and *Philosophy Compass* among others.

Jordan Theriault is pursuing a doctoral degree in psychology at Boston College under the supervision of Dr Liane Young. His research utilizes a neuroimaging approach to explore how moral content is represented in the brain, and specifically, the neural representation of moral objectivity. Additionally, his work explores how moral judgments are adjusted as additional information about persons and situations is encountered. Jordan was born in Halifax, Nova Scotia, and received his B.Sc.H from Queen's University in 2010.

Hannah Tierney is a doctoral candidate in the department of philosophy at the University of Arizona. She has broad philosophical interests, but works mainly on issues of moral responsibility, personal identity, and the self. She is also interested in experimental philosophy and cognitive science. Her publications include "A maneuver around the modified manipulation argument" (*Philosophical Studies*, 2013) and "Keith Lehrer on the basing relation" (*Philosophical Studies*, 2012).

Liane Young is an assistant professor in the department of psychology at Boston College, where she directs the Morality Lab. Her research focuses on the cognitive and neural bases of human moral judgment, including the roles of mental state reasoning and emotional processing. Her research relies on the tools of social psychology and cognitive neuroscience, including functional magnetic resonance imaging, transcranial magnetic stimulation, and the study of patients with cognitive and neural deficits. Young received her BA in philosophy (2004) and her PhD in cognitive psychology (2008) from Harvard University, after which she did postdoctoral work in cognitive neuroscience at MIT's brain and cognitive sciences department.

1

Introduction

Justin Sytsma

Broadly understood, experimental philosophy involves the use of empirical methods to help answer philosophical questions while seriously engaging with the surrounding philosophical literature. More narrowly, experimental philosophy is often described as the use of empirical methods to study intuitions about philosophically interesting scenarios. While I advocate for a broad understanding of experimental philosophy, in general, there is no denying that the narrow conception does a better job of forming a coherent (if selective) whole out of the disparate methods, projects, goals, and motives that animate experimental philosophy.[1] This is especially clear with regard to experimental philosophy of mind, where we find both a central core focused on intuitions about conscious mental states and a periphery that extends out into a wide range of issues.

Articulating experimental philosophy of mind in such a way that it forms a coherent whole might be thought to be especially important, as the unifying threads of the subfield threaten to snap on a broad understanding. One issue is that while it is often difficult to draw a firm dividing line between experimental philosophy and work in the sciences tackling topics of philosophical interest, this difficulty is quite pronounced for experimental philosophy of mind. Brain scientists are often concerned with questions that are of deep interest to philosophers of mind—questions about the nature of the mind, mental states, consciousness, and how these are brought about by, or otherwise relate to, the brain—and they often engage (more or less seriously) with at least parts of the surrounding philosophical literature. The result is that a good deal of scientific work on topics like consciousness, attention, theory of mind, and so on could quite reasonably be included under the label of "experimental philosophy of

mind." Doing so, however, would make it easy to overlook the interesting and important projects that self-proclaimed experimental philosophers have been primarily concerned with. And it is these projects that are the central focus of this volume.

Following the broad/narrow distinction noted above, one way to restrict the domain of experimental philosophy of mind would be to focus on studies investigating people's intuitions. And, in fact, much of what is typically thought of as experimental philosophy of mind can be construed in this way (although it depends somewhat on how one defines "intuitions"). In line with this approach, in the following chapter Jennifer Nado considers recent arguments concerning the use of intuitions as evidence in philosophy. She targets how these arguments relate to claims about intuitions in philosophy of mind, focusing on the use of thought experiments in the literature. For example, Nado discusses Ned Block's (1978) Nation of China thought experiment. Block asks us to imagine the population of China working together to simulate the functioning of a normal human brain. He then argues that according to functionalist accounts of the mind, such a system would have the full spectrum of mental states that you or I have; but, Block claims that our intuitions tell us otherwise—they tell us that the nation of China lacks mental states altogether. Is Block correct in taking his intuition about this scenario to be widespread, though? And, if so, does this intuition constitute good evidence against functionalist theories? As we will see, the work of experimental philosophers casts some light on these questions.

With regard to Block's thought experiment, the intuition at issue is about whether mental states can be correctly ascribed to a particular group agent (the nation of China). And, much of the experimental philosophy of mind literature has focused on intuitions of this sort: Experimental philosophers of mind have been especially concerned with describing and explaining the mental state ascriptions that people make. In fact, the label "experimental philosophy of mind" has primarily been applied to research focusing on questions raised by two seminal articles exploring lay mental state ascriptions—Knobe and Prinz (2008) and Gray et al. (2007). In particular, researchers have explored how lay people classify different mental states and how such classification relates to their moral judgments.[2]

One way to motivate these core issues for experimental philosophy of mind is to note that while psychologists have done a great deal of work on theory of

mind in recent years—work on how people ascribe mental states to themselves and others—this research has largely focused on ascriptions of mental states like beliefs and desires, and their role in reasoning about and predicting agentive behavior. There are another set of mental states, however, that have been at the forefront of philosophical debates about the mind since at least the early modern period—states like feeling pains, seeing colors, hearing sounds, and so on. Philosophers have often taken such *subjective experiences*, to employ the terminology used by Sytsma and Machery (2010), to be quite special. These states are thought to be *phenomenally conscious*: In brief, it is thought that in contrast to mental states like beliefs and desires, there is "something it is like" (Nagel 1974) to have subjective experiences, where this is often cashed out in terms of these mental states having distinctive phenomenal qualities (the pain felt, the color seen).

Having noted the common philosophical distinction between phenomenally conscious mental states and other mental states, several experimental philosophers have asked whether nonphilosophers draw this distinction. More carefully, experimental philosophers have explored whether lay people tend to ascribe mental states to entities in ways that are consistent with the philosophical distinction. Responses that are consistent with this distinction are then taken as evidence that lay people have something like the philosophical concept of phenomenal consciousness, while responses that are inconsistent with this distinction are taken to suggest against the hypothesis. For example, Knobe and Prinz (2008) present evidence suggesting that lay people do in fact have the philosophical concept of phenomenal consciousness. In one study, Knobe and Prinz gave participants a series of ten sentences ascribing different mental states to a group agent (Acme Corporation). They found that participants rated each of the five sentences ascribing mental states that philosophers typically take to be phenomenally conscious as sounding less natural than each of the five sentences ascribing mental states that philosophers typically take to be nonphenomenal. Based on these results, Knobe and Prinz argue both that lay people distinguish between mental states that are and are not phenomenally conscious, and that ascriptions of phenomenally conscious mental states depend on more than just functional properties—the composition of the entity matters as well.

Of course, Knobe and Prinz's latter conclusion is germane to Block's Nation of China thought experiment, as Nado notes. Suppose, for the moment, that

Block is correct and that his intuitions about the nation of China are widely shared. What should we conclude from this? If Knobe and Prinz are correct, perhaps not as much as we might have thought—it might be that our reluctance to ascribe phenomenally conscious mental states to group agents is driving the intuition that the nation of China lacks mental states altogether.

In Chapter 3, Wesley Buckwalter and Mark Phelan also consider Block's Nation of China thought experiment. They call on this example to illustrate the common philosophical claim that a necessary condition for an entity having mental states (or having certain sorts of mental states) is that it is made of the right kind of stuff. As we saw above, Knobe and Prinz defend a related claim—they argue that *lay ascriptions* of subjective experiences are sensitive to the type of body that an entity has. Buckwalter and Phelan challenge this *embodiment hypothesis*, extending their previous defense of analytic functionalism as an account of lay mental state ascriptions (Buckwalter and Phelan 2013; Phelan and Buckwalter forthcoming). They present the results of five new experiments that test the willingness of lay people to ascribe a certain type of subjective experience (emotional states) to disembodied ghosts and spirits. The results of their studies suggest against the embodiment hypothesis.

While Buckwalter and Phelan challenge the second conclusion drawn by Knobe and Prinz (2008), another line of research challenges the first conclusion—that lay people have the philosophical concept of phenomenal consciousness. Building off of their (2009) critique of Knobe and Prinz, Sytsma and Machery (2010) argue that a common philosophical claim is mistaken: It is often claimed that the existence of phenomenally conscious mental states is obvious from first-person experience with states like seeing red and feeling pain. If this is correct, however, then we should expect that lay people will tend to treat these states similarly. For example, we would expect them to deny that a simple nonhumanoid robot could be in either state. This is not what Sytsma and Machery found, however: While lay people tend to deny that such a robot feels pain, they tend to affirm that it sees red.

A number of objections have been raised against this work, but the most common is to argue that the phrase "sees red" can be understood in two ways—in an *informational sense* (the entity makes the relevant discriminations between visual stimuli) or in a *phenomenal sense* (the entity is in the corresponding phenomenally conscious mental state).[3] It is then argued that in Sytsma and

Machery's studies, lay participants tend to adopt the informational rather than the phenomenal understanding. For example, an objection of this sort is given by Fiala et al. (2014) in defending their *agency model* of mental state ascriptions—the view that lay mental state ascriptions result from a dual-process cognitive system, where one process focuses on cues such as having facial features, displaying interactive behavior, and moving with distinctive trajectories in determining whether an entity is a suitable target for mental state ascriptions (see Arico et al. 2011; Fiala et al. 2011).

Fiala, Arico, and Nichols present evidence suggesting that while lay people tend to understand phrases like "sees red" in the phenomenal sense, participants in Sytsma and Machery's studies answered that the robot saw red because this was the only way that they could convey that the robot performed relevant information-processing involving color detection. A response to Fiala and colleagues' version of the informational/phenomenal objection is offered in Sytsma (2014). In addition, in Sytsma (2010b) I offered a general response to this objection based on the hypothesis that lay people do not generally hold the view of colors that the objection presupposes—they do not typically associate colors with mental states, but take them to be mind-independent qualities of things in the world. I then presented empirical results supporting this hypothesis, as well as preliminary evidence that lay people also tend to adopt a similar view of pains—they tend to think of pains as being features of body parts rather than mental states.

In Chapter 4, Kevin Reuter, Dustin Phillips, and Justin Sytsma build off of this work, as well as the related results in Reuter (2011), to argue against a common view of the ordinary concept of pain—that pains are private (no one can feel anybody else's pain), subjective (there are no unfelt pains), and that pain hallucinations are impossible. While Sytsma (2010b) and Reuter (2011) present evidence that lay people do not tend to hold that pains are necessarily private or subjective, they did not investigate lay views concerning pain hallucinations. Reuter and colleagues fill that gap in their contribution to this volume, presenting the results of a series of new studies that indicate that the ordinary concept of pain does in fact allow for the possibility of pain hallucinations.

In addition to the two conclusions just discussed, Knobe and Prinz (2008) also argue that ascriptions of phenomenally conscious mental states are used

to do more than to reason about and predict agentive behavior—they hold that they also play a role in facilitating moral judgments. A similar hypothesis is found in the work of Robbins and Jack (2006) and Gray et al. (2007). The basic idea is that lay people distinguish between subjective experiences and other mental states, and that ascriptions of the former are linked to judgments of *moral patiency* (that morally right or wrong action can be done to the entity), while ascriptions of the latter are linked to judgments of *moral agency* (that an entity is capable of morally right or wrong action). The new research presented in Chapters 5 and 6 expand on this work linking mental state ascriptions with moral judgments.

In Chapter 5, Jordan Theriault and Liane Young review the literature on mental state ascriptions and moral judgments, interpreting it from the perspective of Dennett's discussion of the *intentional stance* and the *physical stance* (Dennett 1987). Following Robbins and Jack (2006), Theriault and Young argue that a further stance is needed—the *phenomenal stance*—that is involved in ascriptions of subjective experiences. Specifically, Theriault and Young argue that the intentional stance alone is unable to explain the sense of moral concern that humans feel for some entities, and they point to the phenomenal stance as being critical to such attributions of moral standing. At the same time, following the work of Sytsma and Machery (2012b), Theriault and Young suggest that moral concern might not be fully explicable in terms of the phenomenal stance either, with the intentional stance also playing a role in attributions of moral standing.

In Chapter 6, Anthony Jack, Philip Robbins, Jared Friedman, and Chris Meyers build off of the distinction between the intentional stance and the phenomenal stance presented in Robbins and Jack (2006) to argue that moral cognition involves two types of psychological processes. Unlike the view put forward by Joshua Greene (e.g., Greene 2007), however, they don't characterize these processes in terms of the contrast between reason and passion; instead, Jack and colleagues characterize the processes in terms of a contrast between two types of reasoning—between a cognitive mode that evolved for interacting with inanimate objects and a cognitive mode that evolved for interacting with conscious agents. While Jack and colleagues associate the former with Dennett's physical stance, they associate the latter with the phenomenal stance. In line with the speculation of Theriault and Young, the intentional stance is then

thought of as involving both cognitive modes. As discussed in the chapter, this model receives support from recent evidence from cognitive neuroscience, as well as from five new experiments reported by Jack and colleagues.

Of course, philosophy of mind is a broad area touching on a wide array of philosophical topics. Not surprisingly, the opportunities for bringing empirical methods to bear on issues in philosophy of mind are correspondingly large; they are certainly not exhausted by issues related to mental state ascriptions and moral judgments alone. This volume closes with a pair of examples—two fascinating contributions to the literature that focus on areas beyond the core issues in experimental philosophy of mind discussed above.

In Chapter 7, Hannah Tierney, Chris Howard, Victor Kumar, Trevor Kvaran, and Shaun Nichols explore the intuitions of lay people with regard to personal identity. They note that while it often makes sense to assume conceptual monism in our philosophical theorizing, there is no guarantee that this assumption is correct in any given case. This is a point that is relevant to many areas of experimental philosophy of mind, including work on mental state ascriptions and moral judgments. The risk of assuming conceptual monism is amply illustrated by Tierney and colleagues for the case of personal identity. They present a range of evidence indicating that lay judgments about the persistence of persons follow (at least) two different criteria—one concept of personal identity conforming to a psychological criterion, while another conforms to a biological criterion. From this, Tierney and colleagues conclude that pluralism best explains lay intuitions about personal identity; they don't stop there, however, but go on to argue that pluralism about personal identity is also a viable philosophical position.

In Chapter 8, Edouard Machery extends on his previous work (Machery 2005, 2009) arguing for a type of pluralism about concepts: According to the *heterogeneity hypothesis*, concepts don't form a unified kind, but instead split into three types that have little in common—prototypes, exemplars, and causal theories. If the heterogeneity hypothesis is correct, and if certain words lexicalize more than one of these distinct kinds of coreferential concepts (the *polysemy hypothesis*), then we would expect to find cases in which competent speakers are willing to endorse seemingly contradictory sentences because they read those sentences in terms of different coreferential concepts. Expanding on the empirical evidence given by Machery and Seppälä (2010), in this chapter

Machery provides new evidence that this is in fact the case: He presents the results of a series of studies indicating that a substantial proportion of English speakers are willing to endorse a surprising range of seemingly contradictory sentences.

Notes

1 See Sytsma and Livengood (forthcoming) for an introduction to experimental philosophy on the broad conception; see Alexander (2012) for an excellent introduction adopting the narrow conception.
2 For a survey of work on these themes see Sytsma (2010a); see Knobe (2008) and Machery and Sytsma (2011) for short introductions. For further discussion of the connection between mental state attributions and moral judgments, see Phelan and Waytz (2012). Of course, there are exceptions to this characterization of experimental philosophy of mind, including important empirical work by experimental philosophers on other issues in philosophy of mind (see, e.g., Hurlburt and Schwitzgebel 2007; Schwitzgebel 2011; Cohen and Nichols 2010). In fact, as noted below, this volume closes with two new examples of such work.
3 For the informational/phenomenal objection see Huebner (2010, p. 137) and Fiala et al. (2014); for other objections, see Talbot (2012) and the response in Sytsma and Machery (2012a), as well as Peressini (2013), Buckwalter and Phelan (2013), and Sytsma (forthcoming).

References

Alexander, J. (2012), *Experimental Philosophy: An Introduction*. Malden, MA: Polity Press.
Arico, A., Fiala, B., Goldberg, R. F., and Nichols, S. (2011), "The folk psychology of consciousness." *Mind & Language*, 26(3), 327–52.
Block, N. (1978), "Troubles with functionalism," in W. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology* (Minnesota Studies in the Philosophy of Science, vol. IX). Minneapolis: University of Minnesota Press, pp. 261–326.

Buckwalter, W. and Phelan, M. (2013), "Function and feeling machines: A defense of the philosophical conception of subjective experience." *Philosophical Studies*, 166(2), 349–61.

Cohen, J. and Nichols, S. (2010), "Colours, colour relationalism and the deliverances of introspection." *Analysis*, 70(2), 218–28.

Dennett, D. (1987), *The Intentional Stance*. Cambridge: MIT Press.

Fiala, B., Arico, A., and Nichols, S. (2011), "On the psychological origins of dualism: Dual-process cognition and the explanatory gap," in E. Slingerland and M. Collard (eds), *Creating Consilience: Integrating the Sciences and the Humanities*. Oxford: Oxford University Press, pp. 88–110.

Fiala, B., Arico, A., and Nichols, S. (2014), "I, Robot," in E. Machery and E. O'Neill (eds), *Current Controversies in Experimental Philosophy*. New York: Routledge.

Gray, H. M., Gray, K., and Wegner, D. M. (2007), "Dimensions of mind perception." *Science*, 315, 619.

Greene, J. D. (2007), "The secret joke of Kant's soul," in W. Sinnott-Armstrong (ed.), *Moral Psychology, Vol. 3: The Neuroscience of Morality*. Cambridge, MA: MIT Press, pp. 35–79.

Huebner, B. (2010), "Commonsense concepts of phenomenal consciousness: Does anyone *care* about functional zombies?" *Phenomenology and the Cognitive Sciences,* 9, 133–55.

Hurlburt, R. and Schwitzgebel, E. (2007), *Describing Inner Experience? Proponent Meets Skeptic*. Cambridge: MIT Press.

Knobe, J. (2008), "Can a robot, an insect or God be aware?" *Scientific American Mind*, 19, 68–71.

Knobe, J. and Prinz, J. (2008), "Intuitions about consciousness: Experimental studies." *Phenomenology and the Cognitive Sciences*, 7, 67–83.

Machery, E. (2005), "Concepts are not a natural kind." *Philosophy of Science*, 72, 444–67.

—(2009), *Doing without Concepts*. New York: Oxford University Press.

Machery, E. and Seppälä, S. (2010), "Against hybrid theories of concepts." *Anthropology & Philosophy*, 10, 97–125.

Machery, E. and Sytsma, J. (2011), "Robot Pains and Corporate Feelings." *The Philosophers' Magazine*, 1st Quarter, 78–82.

Nagel, T. (1974), "What is it like to be a bat?" *The Philosophical Review*, 83, 435–50.

Peressini, A. (2013), "Blurring two conceptions of subjective experience: Folk versus philosophical phenomenology." *Philosophical Psychology*, doi:10.1080/09515089.2013.793150

Phelan, M. and Buckwalter, W. (forthcoming), "Analytic functionalism and mental state attribution." *Philosophical Topics*.

Phelan, M. and Waytz, A. (2012), "The Moral Cognition/Consciousness Connection." *Review of Philosophy and Psychology*, 3(3), 293–301.

Reuter, K. (2011), "Distinguishing the appearance from the reality of pain." *Journal of Consciousness Studies*, 18(9–10), 94–109.

Robbins, P. and Jack, A. I. (2006), "The phenomenal stance." *Philosophical Studies*, 127, 59–85.

Schwitzgebel, E. (2011), *Perplexities of Consciousness*. Cambridge: MIT Press.

Sytsma, J. (2010a), "Folk psychology and phenomenal consciousness." *Philosophy Compass*, 5/8, 700–11.

—(2010b), "Dennett's theory of the folk theory of consciousness." *Journal of Consciousness Studies*, 17(3–4), 107–30.

—(2014), "The Robots of the Dawn of Experimental Philosophy of Mind," in E. Machery and E. O'Neill (eds), *Current Controversies in Experimental Philosophy*. New York: Routledge.

—(forthcoming), "Revisiting the valence account." *Philosophical Topics*.

Sytsma, J. and Livengood, J. (forthcoming), *The New Experimental Philosophy: An Introduction and Guide*. Peterborough, ON: Broadview Press.

Sytsma, J. and Machery, E. (2009), "How to study folk intuitions about phenomenal consciousness." *Philosophical Psychology*, 22(1), 21–35.

—(2010), "Two conceptions of subjective experience." *Philosophical Studies*, 151(2), 299–327.

—(2012a), "On the relevance of folk intuitions: A reply to Talbot," *Consciousness and Cognition*, 21, 654–60.

—(2012b), "The two sources of moral standing." *Review of Philosophy and Psychology*, 3(3), 303–24.

Talbot, B. (2012), "The irrelevance of folk intuitions to the 'hard problem' of consciousness," *Consciousness and Cognition*, 21(2), 644–50.

2

The Role of Intuition

Jennifer Nado

Since its inception, experimental philosophy has been bound up with methodological questions regarding the status of intuition. Several of the most well-known early experimental studies attracted attention due to their criticisms of "traditional" intuition-based argumentation, in which a philosopher takes intuitive reactions to an imagined case to serve as evidence for or against philosophical claims. When, for example, Weinberg et al. (2001) and Machery et al. (2004) found cross-cultural variation in responses to thought experiments, they took this to provide a serious challenge to the default assumption of a "shared" set of intuitions on which they claimed traditional methodology rests.

Such "negative" projects, of course, reflect only a part of the picture. Over its brief history, experimental philosophy has developed into a cross-disciplinary subfield whose practitioners pursue diverse aims. Some experimental philosophers conceive of their project as involving systematic empirical study of intuitions *qua* intuitions, of the psychological processes that produce them, and of the conceptual frameworks that they reflect. Some view themselves as pursuing a new, more empirical approach to resolving standard philosophical questions. However, even these nonnegative projects raise serious methodological questions. Their empirical methodology is itself a challenge to the image of the philosopher's task as a fundamentally armchair-bound exercise; they suggest that intuitive judgments have a different role to play than perhaps previously thought.

Experimental philosophy in all its forms, then, invites us to reflect on the role of intuition in philosophical methodology. Do the intuitive judgments of nonphilosophers provide just as much evidence as those of philosophers?

Do they perhaps provide better evidence? Does intuition provide any evidence whatsoever, or does the empirical work emerging from experimental philosophy simply show intuitions to be hopelessly biased and inconstant? If intuition does have some evidential worth, how much? How is this evidence to be properly elicited and employed?

Ultimately, our understanding of the implications of experimental philosophy depends heavily on our understanding of the role of intuition in philosophy, and vice versa. It's unsurprising, then, that the rise of experimental philosophy has been paralleled by the emergence of an extensive literature on the evidential value of intuition. These debates are deeply relevant to an assessment of the scope and impact of experimental philosophy—indeed, many of the debates explicitly target experimental findings, or even the very cogency of using experimental methods to approach philosophical problems.

With that in mind, the aim of this chapter is to introduce and taxonomize several recent arguments regarding the evidential status of intuitions in philosophy.[1] Since the particular focus of this volume is experimental philosophy of mind, an important goal will be to show how such arguments might specifically apply to intuitions as used in the philosophy of mind, as well as how experimental philosophy in the area of mind might engage with the discussed arguments.

Philosophy of mind deals especially heavily in intuition—thought experiments regarding zombies, Chinese rooms, and color-deprived neuroscientists are the coin of the realm.[2] There is, therefore, deep potential for experimental philosophy to have dramatic effects on our conception of this subfield. In the next section, we'll look at an example from philosophy of mind in order to introduce what I call the "traditional" use of intuitions in philosophy. We'll also look at how experimental philosophy has uncovered data which may make us rethink this traditional use. This will serve as an opportunity to get clearer on the notion of intuition, and on the potential questions regarding its evidential status. After this introductory case study, we'll begin an investigation of some of the major forms of argument found in the intuition literature.

1 Intuitions in philosophy of mind: A case study

In his 1978 paper "Troubles with Functionalism," Ned Block proposed a compelling thought experiment. Imagine that we were to convince each citizen

of China to implement the functional properties of a single neuron, using two-way radios for communication. Suppose we were to use this system to produce, for a single hour, a functional duplicate of your brain. Would such a system be a mind? A functionalist theory of mind would entail that it is. Block claimed that cases of this variety "embarrass all versions of functionalism in that they indicate functionalism is guilty of liberalism—classifying systems that lack mentality as having mentality" (Block 1978, p. 277). The China brain case is, at least *prima facie*, a counterexample to functionalism—for, "there is *prima facie* doubt whether it has any mental states at all—especially whether it has what philosophers have variously called 'qualitative states,' 'raw feels,' or 'immediate phenomenological qualities'" (Block 1978, p. 281).

This "prima facie doubt" involves an intuition, as Block later clarifies. Upon considering the imagined case, we seem drawn to the claim that the China brain would not be a mind—we simply feel that it must be so, though we may not immediately be capable of explaining why. The details of this phenomenon are highly debatable. Intuition has been variously conceived as a type of judgment, a type of belief, an inclination to believe, a *sui generis* propositional attitude, and so on. However, we can point to at least two features to which all sides seem to agree. First, intuitions are marked by an absence of conscious reasoning. Second, they involve a distinctive phenomenological component; this is often described in terms of "seeming." Leaving aside the deeper questions of their nature, at least for the time being, we can give a "first pass" characterization of intuitions as follows: they are states in which a certain proposition seems to be true in the absence of awareness of reasoning.

Intuitions are standardly given at least some degree of evidential weight in philosophical argumentation. That is to say, the propositions which intuition endorses are generally taken to be supported by the existence of said intuitions. The degree of support thus provided is not always clear. In some contexts, certain propositions are treated as nearly sacrosanct solely due to their intuitive appeal—the Gettier case comes to mind.[3] In other cases, it's recognized that intuition may err, and that intuitive premises ought ideally to be supported by further argumentation. Block's case falls into the latter category, at least in its original statement—after introducing the case, Block says that he aims to show that the intuition in fact rests on a rational basis.

Nonetheless, in Block's case (as in others), it still seems as though intuition plays a crucial role. Block's subsequent argumentation for a "rational basis"

appeals, for instance, to the fact that we cannot conceive of how psychology in its current form could possibly explain qualia. This is as much an intuition-based argument as is the China brain thought experiment itself. Block concedes that the arguments here are not decisive, but he does take them to shift the argumentative weight against functionalism. On the "traditional view," then, intuition provides at least some degree of support for the claim that the China brain is not a mind; it thereby provides at least some reason to suppose that case to be a counterexample to functionalism.

Let's now turn to some recent work in experimental philosophy which has bearing on the status of our China brain intuition. Knobe and Prinz (2008) found that subjects are often quite comfortable ascribing beliefs and desires to "group agents"; however, they tend to be hesitant to ascribe phenomenal consciousness to such entities. Microsoft can desire a merger—but it cannot feel depressed. Knobe and Prinz hypothesize that the physical constitution of the subject is a major determinant of our willingness to ascribe phenomenal consciousness, but that it has less effect on our assessment of nonphenomenal mental states; in particular, we seem to resist ascribing phenomenal consciousness to subjects which lack a unified physical body. If this hypothesis is right, it may explain our hesitation to grant the China brain a mind.

Now, Knobe and Prinz themselves had no iconoclastic aims when performing their studies. But at least one similar study has shown that this preference for unified physical bodies may be deeply contingent. Huebner et al. (2010) found evidence that the resistance to ascribing consciousness to group agents may be culturally local—subjects in Hong Kong showed much less reluctance to so ascribe. If these studies are taken at face value, they at least raise the possibility that our reluctance to countenance group agents is nothing more than a quirk of our own psychology—and potentially a culturally mediated quirk, at that.

The empirical work here reflects two potential routes through which experimental philosophy can give us new perspectives on classic puzzles in the philosophy of mind. First, they can propose psychological explanations for intuitions that may otherwise have seemed fairly "brute"—for instance, by proposing that we have a cognitive tendency to invoke physical criteria in our assessment of whether something is a potential experiencer. Second, they can challenge the evidential status of those intuitions, through pointing to factors

such as cultural background which may affect intuition in inappropriate ways. This latter leads us to the first major argument type found in the intuition debates—what I will call the "variation argument."

2 The variation argument

'Variation arguments' are those which appeal to the phenomenon of variation in intuition, either within or across subjects, to cast doubt on the evidential status of intuition. Arguments based on variation share the following very simple argumentative structure. First, it is claimed that if intuitions about some philosophical notion N vary as a function of a certain feature F, then the project of using such intuitions to characterize the nature of N is misguided. Second, empirical evidence is offered in support of the claim that intuitions do in fact vary as a function of feature F. It is then concluded that the project of using intuitions to characterize the nature of N is misguided. This style of argument is, of course, one of the most well-known manifestations of "negative" experimental philosophy.

Perhaps the most common feature F is cultural background of the subject. In a paradigmatic study of this kind, Weinberg et al. (2001) presented subjects of either Western or East Asian background with vignettes describing a Gettier scenario. Though the Western subjects mirrored Western philosophers in judging that Gettier cases were not cases of knowledge, East Asian subjects disagreed. Similar results were found with "truetemp" cases—Western subjects' intuitions mirrored the intuitions of Western philosophers, but East Asian subjects' intuitions did not. The claim is that, unless we are willing to embrace epistemic relativism, we must reject one group's intuitions as false; however, in the absence of some sort of error theory, it would appear to be nothing but naïve ethnocentrism to doggedly hold that the Western responses are the correct ones. Weinberg et al. conclude that intuition should not be the basis for normative epistemological claims.

Cultural variation arguments are only one species of the variation argument type; variation arguments can also be constructed by substituting other features for F in the argument schema described above. In order for the first premise of the argument to be plausible, of course, the feature F which produces variation

must be irrelevant, in the sense that its variation in the cases under evaluation does not plausibly imply variation in the truth value of the hypothesis at hand. Imagine that a study shows that intuitions about whether a case counts as knowledge vary as a function of whether the person in the described case is in possession of a reliable belief-forming mechanism. No variation argument looms here, for the feature F causing the variation in intuition—possession of a reliable belief-forming mechanism—is relevant. It is plausible that the presence of this feature could affect the truth value of an ascription of knowledge.

On the other hand, imagine a study shows that intuitions about whether a case counts as knowledge vary as a function of the order in which cases are presented. A variation argument is now on the horizon, for case order is arguably irrelevant.[4] If nothing but the case order is changed, and yet our intuitions vary, then we have prima facie evidence that at least some of those intuitions are tracking something other than the truth. Swain et al. (2008) have, in fact, run just such a study; they found that subjects were much less willing to ascribe knowledge to the subject in a Truetemp case when they were first presented with a clear case of knowledge. Conversely, subjects were more likely to ascribe knowledge when they were first presented with a clear case of nonknowledge.

There is, of course, disagreement over the significance of this sort of data. Sosa (2009), for instance, has argued that the variation in Gettier intuitions observed in Weinberg et al. (2001) might be explained by a tendency for the two cultural groups to interpret the vignettes in different ways.

> It is not clear exactly what question the subjects disagree about. In each case, the question would be of the form: 'Would anyone who satisfied condition C with regard to proposition <p> know that p or only believe it?' It is hearing or reading a description of the example that enables the subjects to fill in the relevant C and <p>. But can we be sure that they end up with exactly the same C and <p>? (Sosa 2009, p. 107)

It is possible that the cultural differences between the subjects lead them to fill in details not explicitly specified in the vignettes in different ways—perhaps the two groups vary in some crucial background beliefs regarding fake barns, Ford cars, and the like.

This suggestion is of course an empirical hypothesis, but it is not without plausibility. In the context of philosophy of mind we can easily imagine, for

example, intuitions on Mary's room (Jackson 1982) differing as a result of certain background beliefs about the contents and scope of neuroscience. The cultural variation studies performed to date have not typically attempted to test for the presence of such differing background assumptions; this is a genuine weakness in the variation argument. It is, however, a defect that might be remedied by careful empirical work in the future—and this is therefore one area where experimental philosophers can continue to contribute to the intuition debate. In the meantime, the amassed variation studies should still cause the traditionalist a fair bit of concern.[5]

Sosa provides a second, similar concern for the Weinberg et al. findings—one that reflects a very common immediate reaction to the data. It is possible that the Western and East Asian groups have a purely verbal disagreement; "knowledge" might express different concepts for the two groups. There is no genuine disagreement, after all, if Westerners ascribe $knowledge_1$ to a subject in a Gettier case while East Asians withhold an attribution of $knowledge_2$.

One trouble with this suggestion, however, is that it leaves us with a serious normative question—which of the two sorts of epistemic states *ought* we to pursue? Many epistemologists take knowledge to be the primary goal of epistemic activity. However, if the above "divergent concept" case obtains, we're left with a puzzle—is $knowledge_1$ or $knowledge_2$ the primary goal of epistemic activity? There's a clear possibility that the Western tendency to value $knowledge_1$ over the East Asian $knowledge_2$ might amount to nothing more than cultural preference.[6] Of course, there are other responses one might give here. One possibility is to embrace some form of epistemic relativism—though this has not been a popular strategy. Sosa's response, by contrast, is pluralist; he argues that both concepts might express valuable epistemic goals, and that our preference for the one should not preclude our valuing the other.

Stich (2009) has rightly noted that Sosa's response becomes much less plausible when one considers extreme cases of moral disagreement. I would add that Sosa's response is also problematic for nonnormative, philosophically relevant phenomena—like those which form the subject matter of philosophy of mind. If "consciousness" picks out $consciousness_1$ for Westerners and $consciousness_2$ for East Asians, it is not at all obvious that our response should be to simply embrace both concepts as equally legitimate and valuable. In the face of divergent concepts, we must determine which notion deserves a place

in a satisfactory theory of the mind. The answer, of course, *might* be both—say, if the notions play different explanatory roles, and each of those roles is found to have theoretical import. But the point is that a theoretical question remains, even after we've successfully determined which states count as $conscious_1$ and which as $conscious_2$.

3 The calibration argument

A second form of broadly anti-intuition argument appeals to the fact that intuition appears to be resistant to calibration. The original form of this argument, found in Cummins (1998), takes as a starting point the common suggestion that the role intuition plays in philosophical inquiry is analogous to the role that observation plays in the sciences. In the sciences, of course, great care is taken to ensure that the observational procedures employed are accurate. And, Cummins claims, the typical way to do this is through calibration. Before we deem a scientific procedure—say, the use of a new telescope—to be an acceptable source of evidence, we apply that procedure to something whose properties are already known. If the output of the new procedure matches our independent knowledge, this bodes well for the procedure's accuracy; calibration, then, increases our confidence in a procedure's reliability.

Unfortunately, according to Cummins, the analogy between intuition and observation breaks down at this point, since philosophers do not attempt to calibrate intuition. This is no fault of the philosophers; intuition is, in many cases, impossible to calibrate. We simply do not have *independent* access to the sorts of facts that intuition purportedly reveals—how, other than via intuition, can we determine for example whether a certain act is morally good? One tempting response is that we infer truth from widespread agreement; the cases we calibrate our intuition on are the uncontroversial cases. This, however, would be like trying to calibrate a telescope by comparing the results it gave to those given by other telescopes of the same make. Though flaws in that particular telescope might be thus detectable, a design flaw common to the group would not.

Cummins does suggest that there are *some* cases in which intuition can be calibrated; we can, for instance, ask what notion of time our best physical

theory demands and then check intuition's deliverances against this notion. Thus, we can calibrate intuition by theory. The catch-22 for the traditionalist, according to Cummins, is that if we were in possession of a well-developed theory which afforded us independent access to the targets of intuition, we would no longer have need for intuition. "Philosophical theory in such good shape is ready to bid the Socratic midwife farewell and strike out on its own in some other department" (Cummins 1998, p. 118).

Even leaving this problem aside, however, the ability to calibrate intuition by theory will be cold comfort unless theory and intuition are found to coincide. In those cases where we have been able to check our more theoretical intuitions against our best theories, intuition does exhibit some degree of error; stock examples include the naïve comprehension principle in set theory and Kant's claim that space is necessarily Euclidian. The degree to which these errors cast doubt on intuition-based methodology is not wholly clear. They do, however, at least raise the possibility that intuition is a rather flawed instrument.

Cummins's ultimate conclusion is that "philosophical intuition is epistemologically useless, since it can be calibrated only when it is not needed" (Cummins 1998, p. 125). Within the context of philosophy of mind, one is reminded here of arguments made by the Churchlands advocating replacement of folk psychology with neuroscience (see Churchland 1981). With the development of modern sciences of mind, one might worry that intuitions simply no longer have anything to contribute; or, at least, that unanswered questions should be approached from within psychology and neuroscience rather than from within philosophy.

However, there are some clear responses that might be given to Cummins's argument as originally formulated. There's room to question the claim that calibration by theory negates the usefulness of intuition. First, Cummins neglects the possibility that theory could provide a check on some subset of intuitions regarding consciousness (for example) while nonetheless remaining silent on other cases where intuition might provide data. That is to say, intuition's applicability might extend further than the theory being used to calibrate it. Second, Cummins suggests that calibration only occurs when two sources' deliverances regarding some proposition P are compared; however, we might easily imagine other forms of calibration. There may be cases where theory only puts certain constraints on accurate intuition—as an example, we

might suppose that intuitions on the nature of mental states must be consistent with the phenomenon of neural plasticity. If they are found to be so consistent, this provides partial support for the reliability of intuition.[7]

Another response, offered by Goldman (2007), claims that calibration against independently validated procedures is simply too strict a requirement on sources of evidence. Intuition is not standardly calibrated; but neither is observation. Calibration is used on telescopes, but not obviously on vision itself. Goldman claims that basic evidential sources like perception, memory, and introspection are in general resistant to calibration. Just as in the case of intuition, we do not have procedures for accessing the relevant facts that do not ultimately rely on the faculties being tested. Yet, we do not reject perception or memory as sources of evidence.

It's plausible that we are justified in employing a basic source of evidence even if we have not performed a thorough assessment of its reliability. This is not, however, to say that such sources are immune to criticism. Goldman maintains that a weaker condition holds—we must not be justified in believing that the evidential source in question is *un*reliable. As mentioned above, there are a number of cases in which intuition has proven to be in error. But of course, we know that perception is fallible as well, and perception's fallibility does not impugn it as an evidential source. What's needed for a successful calibration argument against intuition is some more robust sense in which intuition resists calibration.

Weinberg (2007) and Weinberg et al. (2012) provide a version of the calibration argument which responds to such worries. On this modified calibration argument, the notion of calibration is expanded to involve what Weinberg et al. call "extrapolative" calibration. During the process of extrapolative calibration, we employ theoretical information about the procedure or instrument in question, in addition to external checks against an independent source. This can grant us confidence that the procedure will be reliable even in cases where independent access is not available.

In order to successfully infer from cases where we are able to perform independent checks to cases where we are not, we must have some idea of how the procedure or instrument operates. By examining the output of the procedure both over time and in varied situations, we may be able to detect unexpected and potentially problematic functioning. One test is that of consistency; if a

microscope pointed at a given object produces inconsistent readings at different times, we then have reason to doubt its accuracy. Another way to test for error is to look for features in the environment which might be illegitimately affecting the device.

If we do identify some problematic functioning via some such method, the device may be "rehabilitated"—for instance, it might be outfitted with some means of resisting the interfering factor(s). If rehabilitation is impossible, the device may instead be "restricted"—that is, we may decide to avoid using the instrument under the problematic conditions. If we are not able to successfully calibrate and subsequently rehabilitate or restrict, an epistemic source whose reliability is in doubt will be unable to regain its credibility.

Perception has been calibrated on this extended notion. Perception is indeed fallible, but through examination of its workings, we have come to know quite a bit about the circumstances under which it fails. Further, we are quite good at restricting our use appropriately—we do not, for instance, put much stock in visual perceptions in dark rooms. But it is not nearly as clear that intuition passes this test.

Weinberg (2007) notes that the real trouble for epistemic sources is not *mere* fallibility. Trouble arises only when an epistemic source suffers from unmitigated fallibility, or "hopelessness." There are at least four basic sources of epistemic "hope," the possession of which mitigates fallibility. The first of these is external corroboration, as in Cummins. A second is internal coherence—in the case of intuitions, coherence both within and across subjects. The third is detectability of margins; that is, we must have some means of identifying the conditions under which the instrument or procedure is likely to err. The final source of epistemic hope is theoretical illumination, in the form of a good theory of the workings of the instrument or procedure, explaining why it works and what has gone wrong when it doesn't.

Weinberg argues that intuition appears to lack much by way of epistemic hope. External corroboration is not forthcoming in domains like ethics; in other domains where we have been able to check intuition against theory, intuition often fares poorly. As for internal coherence, the variation studies discussed in the previous section provide substantial cause for doubt in many cases, and for an even greater number of cases we simply have insufficient data.[8] With regard to the third source of hope, intuition *may* occasionally provide the means for

detection of its own margins—intuitions can be felt more or less strongly, for instance. However, according to Weinberg (2007, p. 335), "this gradation is largely unexplored—and unexploited—by current philosophical practice"; further, strongly intuitive assertions like that of the naïve comprehension principle can turn out to be mistaken. Finally, our degree of theoretical illumination with regard to intuition is minimal; we simply have very little understanding of the causal routes through which intuition operates. To use Weinberg's terminology, intuitions are "introspectively opaque"—their most central feature, as we've seen, is that we have no access to the cognitive processes which produce them. On all these dimensions, Weinberg argues, our current ability to calibrate and rehabilitate/restrict our intuitions appears to be quite low.

Of course, some of these epistemic failings may be remediable; in particular, we may hope to eventually formulate a theory of intuition which is as rich and explanatory as is our current understanding of the workings of vision. Indeed, this is an area where experimental philosophy has much potential to contribute to the debate—after all, many experimental philosophers take the characterization of the psychological mechanisms underlying intuition to be their primary goal. In the area of philosophy of mind, we have a reasonably substantial start on characterizing such mechanisms, both from recent experimental philosophy as well as from the more established psychological literature relating to, for example, folk psychology. Of course, even if experimental philosophy does eventually provide, or help to provide, an adequate theory of intuition, it is still at this point an open question whether such a theory would support or undermine our confidence in our intuitions. We cannot yet successfully calibrate intuition; once we can, we may discover that intuition's flaws are so thorough that a program of rehabilitation/restriction would not be worthwhile.

4 The restriction argument

It is possible to hold that intuition is indispensable to philosophical inquiry while simultaneously arguing that its current usage is overly promiscuous. Michael Devitt, Hilary Kornblith, and Brian Weatherson have each suggested

that our use of intuition should be restricted rather than eliminated. This is not to imply that the anti-intuitionists discussed in the previous sections necessarily aim to reject intuition across the board, though in some cases they have made claims that suggest such a view. Restrictionists differ from the philosophers already discussed in that they in fact offer *positive* accounts of the value of intuition. The restrictions they suggest emerge as direct consequences of those positive views.

For Michael Devitt, the restrictionist position is prompted by his view that intuitions are simply a species of theory-laden, empirical judgments—as opposed to a special sort of a priori insight. Intuitions, according to Devitt (2006), differ from other empirical judgments only in that they are made in the absence of conscious reasoning. We can identify two types of intuitions which play a role in inquiry. The first are the sorts of intuitions by which we identify members of a given kind under investigation; our intuitions that this is an F but this is not. Call this "basic" intuition. The second sort of intuition, which we might call "rich" intuition, provides more general judgments about the Fs identified by the basic intuitions; a typical rich intuition would be something like "belief plays a central role in producing action."

There are two stages to an investigation into the nature of a given kind, whether that investigation is philosophical or scientific. During the first stage, we must identify uncontroversial cases of the kind to be investigated. Often this is done in the absence of any theory of the kind we are interested in; in such a case, basic intuitions are crucial. The best sources of appropriate basic intuitions are those persons who have the most empirical expertise with the kind at hand. In some cases—Devitt uses pains as an example—this may be "the folk." In cases where some scientific theory is available, however, intuitions of the relevant scientists are preferable. This is in sharp contrast to the standard philosophical view, upon which we must take pains to avoid intuitions that have been "contaminated" by theory. Theory-contamination, in Devitt's view, is a virtue rather than a vice.

The second stage of investigation, once we have identified samples of the kind in question, is to examine those samples and determine what is "common and peculiar" to them. Rich intuitions may, at this stage, be a source of hypotheses; but they are in no way necessary. Further, as with basic intuitions, rich intuitions should be trusted only insofar as they reflect strong empirical

expertise with the kind at hand. The best method of investigation is direct, scientific investigation of the kind; and where intuition and experimentation conflict, it is intuition which should be rejected.

Hilary Kornblith's (1998, 2002) account of the proper use of intuition in philosophy is quite similar to Devitt's. Kornblith, like Devitt, rejects the a priori view of intuition and takes the activity of philosophy to be analogous to the investigation of natural kinds in the sciences. The purpose of appealing to intuition in philosophy is to "make salient certain instances of the phenomenon that need to be accounted for . . . much like the rock collector who gathers samples of some interesting kind of stone for the purpose of figuring out what it is that the samples have in common" (Kornblith 1998, p. 134). Kornblith also agrees with Devitt that these identification intuitions are theory-laden, and that the influence of background theory (when that theory is accurate) improves rather than degrades the trustworthiness of intuition. Finally, Kornblith and Devitt both agree that this initial process of identification produces only a rough estimate of the boundaries of a class, and that further theory will in many cases show that some of the initial judgments were mistaken.

Brian Weatherson (2003) formulates his proposal for the role of intuition in the context of a defense of the justified true belief (JTB) model of knowledge. The JTB theory is widely considered to be inadequate due to the intuitiveness of Gettier counterexamples. But why should intuition trump theory in such a case? Weatherson claims that the model according to which we aim for a brute "best fit" with intuition is too crude. Instead, there are at least four separate criteria upon which to judge the success of a philosophical theory, not all of which invoke intuition.

First, it is true that a good philosophical theory should not have too many counterexamples. "While a theory can be reformist, it cannot be revolutionary" (Weatherson 2003, p. 6). Second, the theory cannot have too many undesirable theoretical consequences. To take Weatherson's example, a successful ethical theory should not imply that conspicuousness of suffering is a morally relevant feature. Third, the analysis proposed by the theory ought to be one upon which the concept analyzed turns out to be theoretically significant; *ad hoc* analyses are not successful. Finally, the analysis should be simple. Given that a theory might do better than its rivals on two or three of these measures while doing worse with regard to counterexamples, it seems that there should be at least

some cases where theory trumps intuition. Indeed, on these criteria it seems plausible that "knowledge" might mean justified true belief. Though this analysis falsifies a few of our pretheoretical beliefs, it does well on the other three criteria, and is notably simpler than post-Gettier alternatives.

What are the consequences for experimental philosophy on the restrictionist views just mentioned? Arguably, they leave a significant role for experimental methods. If intuition holds evidential weight in some cases, but not others, then there is a clear need to distinguish the usable intuitions from those that must be abandoned. In some cases, this task may be approached via, for example, appeal to the theoretical criteria outlined by Weatherson. But this does not rule out the possibility that experimental investigation of the psychological mechanisms underlying intuition could significantly contribute to the project, as well. For instance, experimental work might reveal, as discussed in the introduction, that intuitions regarding consciousness in group agents are heavily dependent on a somewhat idiosyncratic and potentially inappropriate bias against agents without a unified physical body. This might provide reason to doubt those particular intuitions, without consequence for, for example, the more basic intuition that an average, non-brain-damaged, adult human would count as conscious.

5 The indispensability argument

We move now from arguments broadly critical of intuition, to arguments broadly in defense of intuition. Perhaps the most prima facie compelling argument supporting intuition is that it is just not possible to do without it. In one form, the argument is that intuition is simply so fundamental to reasoning that to reject all uses of intuition would result in a position of complete skepticism. This argument has been made, for instance, by Bonjour (1998) in his defense of "rational insight." Bonjour claims that in the absence of direct, immediate insight or intuition, no reasoning would be possible. Even performing a simple *modus ponens* inference requires, for example, the ability to directly see that one's premises and conclusion form an instance of the relevant argument form, as well as that the inferential pattern itself is valid. At least some very basic forms of intuition, then, must be counted as epistemically respectable.

A more narrowly focused form of the indispensability argument involves the claim that any radical anti-intuitionist thesis—that is, any thesis that claims that intuitions have no evidential weight—undermines itself. Anti-intuitionists argue that intuition should be removed from our evidential resources; but, it's argued, the premises of the anti-intuitionist's arguments can only be defended by appeal to intuition. We might call this the "self-defeat" variant of the indispensability argument.

One recent well-known instantiation of this argument type can be found in Pust (2000, 2001). Pust's target is the "explanationist" objection to the evidential status of intuition, one statement of which can be found in Harman (1977).[9] Harman claims that the best causal explanation for our having the moral intuitions we do does not advert to the truth of the intuited moral propositions, but instead merely to contingent psychological facts about moral reasoning. Because of this, we are not justified in using moral intuitions as evidence for the truth of moral facts. This argument, Pust claims, rests on an unarticulated premise involving a general "explanationist" criterion of justification, which runs as follows: one is justified in believing only those propositions which either (a) report the occurrence of judgments or observations, or (b) figure in the best explanation of the occurrence of those judgments or observations.

However, Pust claims that this explanationist criterion of justification is ultimately self-undermining, as can be seen by use of an analog of Harman's argument. The explanationist criterion suggests that we should only believe *epistemological* propositions when they either report, or figure into the best explanation of, judgments or observations. But if the best explanation for our moral intuitions does not advert to the truth of the corresponding propositions, then it is likely that the best explanation of our epistemological intuitions does not advert to the truth of the corresponding propositions. So we are not justified in using epistemological intuitions as evidence for the truth of epistemological facts. Insofar as the explanationist criterion of justification itself seems to rest on an epistemological intuition, the argument shows that we are not justified in believing in the truth of the explanationist criterion. Since that criterion was a premise in the argument itself, the argument is self-undermining.

There are two obvious places to question this argument. First, we might question the assumption that the epistemological version of the argument is as compelling as the moral version. There is surely no inconsistency in holding

that moral facts do not factor into the explanation of moral intuitings, but that epistemological facts *do* factor into the explanation of epistemological intuitings. If one could motivate the idea that moral reasoning is, for example, more subject to contingent features of our own psychology than is epistemological reasoning, then Pust's argument falls through. Such an argument might invoke the apparently greater influence of emotion on moral intuition, for instance.

Second, we might question the assumption that the explanationist criterion can only be defended via intuition. There are arguably all sorts of ways in which one could be justified in believing the explanationist criterion; notably, one could be justified in believing it because it follows from one's best epistemological theory. And belief in one's best epistemological theory could be justified because it explains all sorts of things, be they intuitions or otherwise.[10]

Another version of the self-defeat argument can be found in Bealer (1992). Bealer's argument aims at the radical empiricist who holds that one's evidence consists only of one's observations or experiences; to use Bealer's terminology, the radical empiricist wishes to formulate an intuition-free alternative to our "standard justificatory procedure." Bealer objects that the empiricist, in formulating this alternative procedure, violates its ban on intuition-based inquiry. The empiricist must surely make use of basic epistemic terms like "observation," "theory," and "explanation" in formulating her new procedure. But how does the empiricist determine what counts as an observation, as a theory, as explanation, or as justification? These basic epistemic classifications—which Bealer calls "starting points"—are arrived at via intuition, even for the empiricist. The empiricist's alternative procedure, then, inevitably undermines itself.

As before, the empiricist/anti-intuitionist might claim that she has nonintuitive justification for making the epistemic classifications that she does; more plausibly, Bealer suggests that she might claim that although she initially formulated her starting points by use of intuition, she no longer relies on intuition for her current justification. Bealer claims that this leads to a fatal dilemma. Intuitions about starting points are either reliable, or they are not; if they are reliable, then they are eligible to serve as evidence and the empiricist's rejection of them is unwarranted. If they are not reliable, then the starting point judgments that the empiricist initially formulated on the basis

of intuition are prone to error. This error, Bealer claims, will be reflected in the theories that result from those starting points—theories which include the empiricist's epistemological principles.

It's unclear why Bealer thinks that the anti-intuitionist can't claim that we employ our other cognitive resources in order to identify and expunge errors generated by our initial intuitions. There is a massive body of propositions which we are justified in believing; it's plausible that this body is sufficient for the construction of theoretical principles which could lead us to correct errors in our more unreliable classification intuitions. However, even leaving this aside, there is a more fundamental reply that one can make to this sort of self-defeat argument, and indeed to indispensability arguments in general.

The reply focuses on the apparent assumption by advocates of the indispensability argument that intuition is *monolithic*. That is, there appears to be an assumption that intuition forms some sort of unified faculty, such that granting evidential status to any intuition would thereby grant that status to them all. But it's not at all clear that intuition is so unified. The psychological mechanisms that produce, for example, epistemological intuitions are quite plausibly separate from the mechanisms underlying, for example, our use of fundamental logical rules, or our intuitions about mental states. One piece of psychological evidence in favor of such a claim is the apparent dissociability of the cognitive skills related to different types of intuition. Psychopathy provides a *prima facie* instance of selective impairment in moral reasoning, without corresponding impairment in other "intuitive" domains; autism provides the same for reasoning about mental states. Certain patterns of damage to the brain can even cause specific impairment in logical reasoning (Reverberi et al. 2009). Such phenomena are not decisive, of course, but they are suggestive of some degree of psychological heterogeneity in intuition.[11]

If we can in fact make principled, psychologically motivated distinctions between different types of intuitive judgment, then it remains an open question whether the reliability or epistemic respectability of one class of intuitions would have any bearing on the reliability of others. Thus, indispensability arguments lose much of their bite: even if a certain subset of intuitions is shown to be required, either to avoid skepticism or to prevent argumentative self-defeat, this by no means serves as sufficient reason to think that traditional methodology is wholly, or even substantially, in the clear.

Here, again, is a place where experimental philosophy has obvious potential to contribute to the ongoing debates. Insofar as experimental philosophy can help us gain insight into the actual psychological mechanisms underlying intuition, it can help us to determine whether the capacity for intuitive judgment is in fact unified, or whether it is more usefully subdivided into several fairly heterogeneous cognitive mechanisms or processes. This can in turn lead to a more nuanced understanding of the epistemological status of different varieties of intuition.

6 The parity argument

As noted in the discussion of calibration, the idea that the role of intuitions in philosophy parallels the role of perception in the sciences is prima facie attractive. Many philosophers take perception to be an uncontroversially justified, "basic" source of evidence. Could intuition be similarly basic? If it could be shown that intuition's epistemological properties are similar to perception's in some relevant ways, one might be able to thereby defend the use of intuition. Perception's evidential status is often taken to be nonnegotiable; one might be able to argue by parity that we ought to extend the same status to intuition.

A highly relevant epistemological similarity between perception and intuition is the following: both are fallible. Further, both exhibit failures which are not merely occasional or random—in many cases, the failures are systematic. This has not prompted us to abandon perception as an evidential source; Sosa (1998, 2007) argues that it should not present a reason to abandon intuition. Studies in experimental philosophy have shown that our intuitions vary when case order is reversed, or when descriptions are reframed; but, Sosa claims, this is simply analogous to perception's susceptibility to various errors in unfavorable conditions.

> Surely the effects of priming, framing, and other such contextual factors will affect the epistemic status of intuition in general, only in the sort of way that they affect the epistemic status of perceptual observation in general . . . the upshot is that we have to be careful in how we use intuition, not that intuition is useless. (Sosa 2007, p. 9)

As we saw earlier, proponents of the variation and calibration arguments have sometimes suggested that their arguments demonstrate that intuition should be wholly rejected as an evidential source; according to Sosa, however, no such conclusion is warranted. In fact, his statement implies that not even a moderate pessimism is appropriate; after all, perception is a fully respectable evidential source despite its flaws.

Weinberg's discussion of hopelessness, discussed earlier, suggests some ways in which intuition might plausibly differ from perception with regard to the epistemological impact of these sorts of errors. Sosa's suggestion that we simply employ caution only helps if we know what to be careful *for*; intuition's poor scores on Weinberg's four criteria of hopefulness imply that, at present, we are not capable of restricting intuition in appropriate ways. Our assessment of Sosa's argument depends on how deep the similarities between intuition and perception really are. Some progress on this question may be made from the armchair, but it seems clear that a deeper theoretical understanding of the cognitive mechanisms underlying intuitive judgment, such as might be provided by work in experimental philosophy, would be of central import.

We have thus far been pursuing an argument for parity between intuition and perception. There is a variant on this argument, proposed by Williamson (2004, 2007), which proceeds in a somewhat different vein. Williamson notes the indispensability and reliability of our counterfactual reasoning and our commonplace practices of applying concepts in everyday judgment. Williamson then argues that philosophical intuition, rather than being some *sui generis* mental activity, is simply an application of these sorts of basic cognitive capacities. The upshot of this is the same as for the argument for parity with perception—if you grant evidential status to the one, you had better grant evidential status to the other. Anti-intuitionist arguments, Williamson claims, apply equally well to any sort of concept application—we cannot consistently reject philosophical intuition without undermining the general practice of applying concepts. And if anti-intuitionists bite the bullet and reject concept application across the board, their position becomes one of extreme skepticism.

Note that Williamson's claim is more radical than Sosa's. It's not *just* that there is an epistemological parity between ordinary concept application and philosophical intuition, but further that there is *no* principled distinction to

be made between the two—either epistemologically or metaphysically. The traditional conception of intuition as a hyper-rational, unified, prototypically philosophical capacity turns out, upon further investigation, to be oversimplified. There's nothing epistemologically distinct about intuition.

Our assessment of this version of the parity argument will depend on whether a principled distinction can in fact be made between philosophical uses of intuition and similar, everyday forms of cognition. At least some effort has been made to so distinguish—see Weinberg (2007) for an attempt along these lines. But again, this is an area where the needed arguments will plausibly deeply involve empirical understanding of philosophical cognition. At the risk of resembling a broken record, experimental philosophers have the potential to take up the task.

7 The constitutivity argument

Constitutivity arguments are characterized by their claim that intuitions must necessarily reveal truths (at least, in suitably good cognitive conditions), due to the existence of some sort of constitutive tie between intuition and meanings or concepts. That intuition is generally reliable simply follows from the existence of this constitutive relation. This is not to suggest infallibility; it is agreed on all sides that we may still err, if we are inattentive or if we do not reflect appropriately. But these errors must lie at the level of performance, rather than the level of competence.

Bealer's version of this form of argument involves the assertion that intuition has a "strong modal tie" to the truth; and further, that the existence of this modal tie implies that philosophy is both autonomous from and authoritative over the sciences. The most compelling argument Bealer offers for the existence of a strong modal tie between rational intuition and the truth is his "multigon" example, which proceeds as follows. Suppose a woman introduces through use (as opposed to via stipulation) a new term, "multigon." She applies this term to pentagons, hexagons, and so forth, but she has not yet applied it to, nor withheld it from, triangles or rectangles. Suppose she now considers whether triangles and rectangles are in fact multigons. Supposing that she is suitably intelligent, attentive, and so forth, and that her term "multigon"

expresses a definite concept that she determinately possesses, Bealer claims that it's plausible that the woman will judge that triangles and rectangles are multigons if and only if the property of being a multigon is identical to the property of being a polygon, and she will judge that triangles and rectangles are not multigons if and only if the property of being a multigon is identical to the property of being a polygon with five or more sides. Thus, the woman's intuitions have a strong modal tie to the truth.

The argument crucially rests on the idea that the woman *determinately* possesses the concept expressed by "multigon." "Determinate" concept possession is a term introduced by Bealer, and is to be contrasted with "nominal" concept possession. We possess a concept nominally if we are able to have propositional attitudes toward propositions whose contents contain that concept. We possess a concept determinately if we possess it nominally, and in addition do not possess it "with misunderstanding or incomplete understanding" (Bealer 2000, p. 11).

At this point, however, the account becomes unsatisfying. By the definition of "determinate" concept possession given, the multigon example merely demonstrates that if we possess a concept without misunderstanding or incomplete understanding, our intuitions will necessarily be generally true. On what I take to be the standard interpretation of the word "misunderstanding," this seems to say that if our possession of a concept involves no falsehood, our intuitions about that concept will necessarily be generally true. But the critic of intuition is not going to lose much sleep over the fact that there is a strong modal tie between nonfalsity and truth.

For his account to have any bite, Bealer needs to show that we do possess philosophically relevant concepts determinately. A step toward such an argument is given by Bealer's claim that philosophical terms are "semantically stable"—that knowledge of their conditions of application does not require any contingent knowledge about the speaker's external environment (in contrast to terms like "water," whose meaning depends on the nature of the watery stuff in one's environment). Since no empirical knowledge is required in order to possess these concepts determinately, there is no barrier to determinate, a priori philosophical understanding—and, "intuitively, it is at least possible for most of the central concepts of the a priori disciplines to be possessed determinately by some cognitive agent or other" (Bealer 2000, p. 12).

The question of whether philosophical terms do in fact possess semantic stability is, to my mind, open to debate. But more importantly, there has still been no real positive explanation of the "strong modal tie." If Bealer is right, then we do not need knowledge of the contingent features of the external world in order to understand when something counts as, for example, a case of knowledge. But surely this is insufficient to guarantee the necessary reliability of our intuitions about knowledge. After all, presumably no knowledge of the contingent features of the external world is required in order to come to know the truths of mathematics; but this fact alone does not guarantee that we are reliable at mathematical reasoning. Nor does that fact explain whatever reliability our mathematical capacities may possess (indeed, our access to mathematical truths remains deeply puzzling).

Though Bealer's account, I would argue, fails to explain the proposed necessary reliability of intuition, such an explanation might be derived from an account of the nature of meaning. A meaning-based approach to the a priori accessibility of philosophical knowledge has, in fact, been offered by Frank Jackson in his book *From Metaphysics to Ethics*. Jackson argues that philosophy requires "serious metaphysics." Serious metaphysicians work with a limited ontology—for the physicalist, this ontology should more or less consist of the properties, objects, relations and so forth employed by completed physics. Most philosophically interesting terms will not explicitly occur in the language describing the fundamental ontology. However, the existence of "higher-level" entities may be entailed by the basic ontology, if the higher-level entities are supervenient on the lower-level entities. Being entailed in this way is both necessary and sufficient for an entity's inclusion in the serious metaphysician's implied ontology. Jackson calls this principle "entry by entailment."

For each entity described in a high-level vocabulary, the metaphysician must either "locate" that entity by showing how its existence is entailed by facts described in the physical vocabulary of her basic ontology, or she must eliminate the higher-level entity from her ontology entirely. Here is where conceptual analysis is required; for, according to Jackson, "conceptual analysis is the very business of addressing when and whether a story told in one vocabulary is made true by one told in some allegedly more fundamental vocabulary" (Jackson 1998, p. 28). Conceptual analysis, by examining folk

intuitions about the extensions of terms in possible scenarios, reveals the implicit understanding of the folk and uncovers the meanings of their terms, and thus the entities that will have a place in our implied ontology.

This is a genuinely explanatory constitutive approach; the fundamental idea is that the folk possess mental theories or descriptions which determine the meanings of their words. Given that one's intuitions reflect these theories, then, one's intuitions reflect facts about the meanings of one's terms.[12] Therefore, our intuitions about hypothetical cases will necessarily be by-and-large true. One potential trouble with this version of the constitutivity strategy, however, is that the divisions effected by folk theory may not be important divisions from the standpoint of scientific theory—depending on the theorist's goals, the "implied ontology" Jackson's method will generate may not be an explanatorily useful one. It is open to the anti-intuitionist to claim that philosophy should be concerned with phenomena that, for instance, aid in explanation and prediction or factor into laws; and further, that there is no reason to think that folk theory will necessarily reflect categorizations that play those theoretical roles.

Alvin Goldman's (2007) version of the constitutivity argument, by contrast, avoids this problem. His strategy is to claim that the truths revealed by intuition are simply truths about one's concepts, in the psychological sense of concept—and further, that a primary aim of philosophy is to uncover these psychological truths. On Goldman's account, possession of a concept is to be understood as possession of a psychological structure underlying one's use of a given natural-language term. Given this characterization of concepts, it seems to simply follow that possession of a concept which underlies a natural language term "F" will involve a disposition to judge "x is an F" when and only when x satisfies that concept. *What it is* to have a given concept is to have a psychological structure which disposes one to make categorizations in accordance with the content of that concept.

As with Jackson's account, this version of the constitutivity approach is genuinely explanatory; the necessary reliability of intuition is grounded in the very nature of concept possession. However, this story only really explains why, for example, *my* intuitions are reliable indicators of facts about *my* individual, psychological concepts. We may be able to move from facts about individual concepts to facts about shared concepts or word meanings, if the members of

our community have concepts which are substantially similar. Where there is substantial disagreement among members of the community, however, even this may not be possible.

Goldman acknowledges this limitation, and his account thus expresses a very modest assessment of the role of intuition. On the other hand, the account does imply that intuitions are genuinely evidential, in that they grant reliable access to truths about our personal psychological concepts; further, this evidential role is not undermined by phenomena like cultural variation. Finally, truths about personal psychological concepts are plausibly explanatorily important, particularly if one's project is avowedly psychological; Goldman will not, therefore, face the problem just raised for Jackson's implied ontology.

Goldman claims that viewing philosophical analysis as targeting psychological concepts provides justification for a good portion of our actual philosophical practices. For instance, it explains why philosophers place a high value on pretheoretical intuition—if one's intuition is influenced by explicit theory, it will not reflect one's underlying concept. Goldman's viewpoint also squares quite well with views expressed by some of the advocates of the positive approach to experimental philosophy. For instance, Knobe and Nichols (2007) suggest that questions about the workings of the mind, including the conceptual structures underlying philosophical judgment, have been traditionally central to philosophy, and that experimental philosophy follows in that tradition by attempting to reveal interesting psychological facts about our intuitive cognition.

Of course, the psychologistic approach to philosophy does not make sense of certain other philosophical practices—for example, "biting the bullet" when one's theory produces counterintuitive results. Further, if our aim were purely psychological, undesirable theoretical features like inconsistency or ontological promiscuity would not provide a reason to reject an analysis—after all, we should not assume that our personal psychological concepts avoid contradiction or make appropriate use of Occam's razor.

One begins to suspect that current philosophical methodology often reflects a running together of both Goldman's aim of characterizing our concepts as well as a Devitt/Kornblith–type aim of delineating explanatorily useful kinds. In fact, it is entirely consistent to pursue both projects. As such, we should agree with Nichols and Knobe that there is nothing antiphilosophical about the

experimental philosopher's attempt to elucidate the psychology of intuition. In some cases, the two approaches may even complement one another; as Goldman notes in earlier writings on metaphysics, an understanding of psychology can indicate to us that some of our metaphysical distinctions are not objective, but merely reflect innate tendencies to, for example, group perceptual elements together in certain ways (Goldman 1987). Psychological investigation can provide the metaphysical prescriptivist with error theories.

8 Questioning the presuppositions

By way of conclusion, we must discuss one final, very recent argumentative category—one which involves the claim that it is a mistake to focus on intuition in the first place. This broad and diverse argument type, which we might call the *false presupposition* argument, urges that the entire dialectic of recent debates over philosophical methodology has been fundamentally misguided. Though the participants have been arguing over intuition's status as a source of evidence for philosophical theories, the truth of the matter is that philosophy doesn't rest on intuitions in the first place.

In fact, the Williamsonian parity argument briefly discussed earlier can be seen as falling into this category as well. Williamson claims that there is nothing epistemologically distinctive about intuition. He writes that "philosophers might be better off not using the word 'intuition' and its cognates. Their main current function is not to answer questions about the nature of the evidence on offer but to fudge them, by appearing to provide answers without really doing so" (Williamson 2007, p. 220). The very practice of construing philosophical arguments as relying on intuition only leads to an illegitimate "psychologizing" of the evidence—that is, to the false notion that our evidence in philosophy consists solely of psychological facts, rather than facts about, for example, knowledge or consciousness.

A similar argument against psychologization is given by Deutsch (2009). Deutsch claims that variation arguments against the use of intuition make the dubious assumption that philosophers are committed to using claims about intuition as premises in their philosophical arguments. But, Deutsch claims, philosophical arguments do not standardly rest on intuition. Machery et al.

(2004) write as though Kripke's arguments against descriptivism rest on an intuition that "Gödel" refers to Gödel. But this is false. Kripke's argument doesn't rest on the *psychological* fact that someone intuits something, it relies on the *semantic* fact that "Gödel" refers to Gödel. Deutsch's conclusion is that the prevalent misassumption that philosophy depends on intuition undermines much of current experimental philosophy.

Finally, Herman Cappelen (2012) has mounted a book-length argument against the idea that philosophy centrally relies on intuitions. Through examination of the actual text of well-known thought experiments, Cappelen argues that philosophers do not generally base their arguments on appeal to anything that might be construed as having the epistemological features assigned to intuition. When they do use terms like "intuition," philosophers typically only mean to indicate that something is pretheoretically plausible, that it is "in the common ground" of the debate, or that they wish to hedge their claim rather than fully endorse it. Cappelen's final chapter pronounces that experimental philosophy has been "a big mistake." Experimental philosophers, both positive and negative, aim to study intuitions—but philosophers don't employ intuition in their theorizing. Thus, "the project of checking people's intuitions is philosophically pointless" (Cappelen 2012, p. 222). Experimental philosophy, Cappelen suggests, is a bankrupt enterprise.

On the contrary, it seems to me that the implications of such views for experimental philosophy are not quite so bleak. There are two separate threads running through the above arguments; the first involves the cogency of the claim that intuitions are evidence. The authors discussed above emphasize that our evidence does not consist of intuitions—our evidence consists of facts, such as the fact that the Gettier case is not a case of knowledge. Experimental philosophers have had a tendency to "psychologize" the evidence, to the detriment of their arguments. Yet, it's not at all clear why "de-psychologizing" the evidence should make investigations of intuition irrelevant to an assessment of our evidential resources. In the sciences, we might say that our evidence consists of facts rather than observations themselves; yet an assessment of the workings of our perceptual faculties is clearly relevant to an assessment of our evidential situation. Should we discover unexpected biases in vision, our assessment of which facts we can take to be evidence would surely be affected.

The other thread of argument reflects a kind of eliminativist take on intuition. That is, it's suggested that there simply isn't a natural category of mental state corresponding to the term "intuition" as ordinarily used by philosophers; hence, we cannot straightforwardly speak of philosophical judgments or arguments as being "based on intuition." There is something deeply plausible about this. Participants in experimental philosophy and in current debates over philosophical methodology have inherited a conception of the processes and methods underlying philosophical argumentation and judgment that is, to put it bluntly, old. It would be wholly unsurprising if this conception was in fact misleading in some seriously fundamental ways. Indeed, we've already discussed earlier in the chapter how it's quite likely that there's no homogeneous "intuition faculty" underlying philosophical cognition. Perhaps our understanding of philosophical method is fundamentally flawed in yet deeper ways, as the authors above suggest. But the claim that any of this invalidates the very idea of experimental philosophy strikes me as wrong-headed.

Philosophers make judgments, and give arguments. Our current "theory" of such activities invokes a psychological category—intuition—that may well go the way of phlogiston or élan vital. But this in no way shows that there is no psychological story to be told about the reasoning processes that philosophers in fact employ, nor that that story (or more probably, *stories*) would be of no use to philosophy generally. It's possible that our current accounts are fairly far off the mark. There may even be fundamental methodological flaws in current experimental methods—for instance, the untutored judgments of nonphilosophers may turn out to shed no light whatsoever on the reasoning processes involved in professional philosophy (though I find this highly doubtful). Experimental philosophers may need to evolve their methods and their vocabulary, and may even need to rethink the conceptual framework upon which their investigations rest. But this is nothing more than the standard lot of the scientist—or, indeed, of any participant in any academic discipline whatsoever.

Perhaps experimental philosophy can be more usefully conceived of as the study of the psychological processes underlying reasoning about philosophical questions—whether or not that involves intuition. There's every reason to think that such study will continue to advance our understanding of philosophical method. Just as importantly, there's every reason to think

that such study increases our understanding of the mind itself. The links between experimental philosophy and psychology are close indeed, as can be seen by a quick glance through the contributions to this volume. Mental state attribution, moral judgment, agency, concept possession—all are topics investigated by psychologists as well as philosophers. No deep distinction seems likely to be forthcoming; nor would such a distinction obviously be desirable. Experimental philosophy is a young field, and numerous methodological and conceptual questions regarding its nature and its relation to traditional methodology remain to be explored. Answering these questions will be no easy task; but then, nothing in philosophy ever is.

Notes

1 The taxonomy I propose in this chapter is far from exhaustive; I do, however, hope to have covered the most prominent views in the recent literature. There are a few important responses to specific anti-intuitionist arguments which do not fit into my proposed taxonomy. I have not, for example, proposed an argument type for Sosa's suggestion that the findings of Weinberg et al. (2001) can be explained by a difference in the propositions the subjects are entertaining rather than by disagreement over a particular proposition (see Sosa 2009). Specific responses of this sort will instead be discussed during exposition of the arguments they respond to, where appropriate.

2 These are all contemporary examples, but there are also earlier examples of mind-related thought experiments—for instance, Molyneux's problem (would a blind man, upon restoration of his sight, recognize shapes on this basis of previous tactile acquaintance?), Hume's claim that it is possible to have an idea of a "missing" shade of blue which one has never perceived, or Leibniz's use of a mill analogy to argue against the possibility of a mechanical explanation for perception.

3 Though, as will be discussed later, Weatherson (2003) provides a rare case where the Gettier intuition is questioned.

4 Though this claim might be contested by epistemic contextualists. See Swain et al. (2008) for discussion of this point.

5 It is worth mentioning that the "differing background beliefs" hypothesis is much less plausible for explaining away framing effect and order effect findings. Sosa in fact has a separate reply to the Swain et al. order effect findings. It is a

reply which draws an analogy between intuition and perception, and is a good example of what I call a "parity" argument. Parity arguments will be discussed in Section 2.2.

6 Of course, it may well be the case that knowledge$_1$ *is* a better epistemic goal than knowledge$_2$; the point is that this thesis cannot be defended solely via intuitions about what falls under the term "knowledge."

7 Thanks to Jonathan Livengood for bringing these points to my attention.

8 Of course, there is also substantial agreement, particularly on very basic cases such as "murder is wrong." However, it's plausible that the requirements on consistency here are rather high. Take as an example a certain make of thermometer which has been found to be quite consistent on temperatures between 10°C and 40°C, but which is fairly inconsistent in its readings when exposed to very high or low temperatures. Insofar as we need temperature readings at those ranges for whatever intellectual enterprise we are engaged in, use of that thermometer will be problematic. There's an analogy here with intuition; there may be widespread agreement on core cases, but it is often the "outlying," unusual cases that decide between philosophical theories. All plausible metaethical views entail that murder is wrong; however, they disagree on more subtle cases. Insofar as intuition is inconsistent in such cases, this is a challenge to intuition's "hopefulness."

9 I have not included "explanationist" arguments in my taxonomy of the primary arguments against intuition, simply because explanationist arguments have not been prevalent in the intuition literature over the last 15 or so years. However, Weinberg's "theoretical illumination" criterion for hopefulness perhaps reflects something of the spirit of the objection—the natural reading of his view is that philosophical use of intuition will be potentially validated if it turns out that there is a causal pathway between intuitions and the facts they purport to reveal, paralleling the case of vision, providing an explanation of why we have the intuitions that we do—and, further, if that explanation makes plausible the hypothesis that intuition generally reflects truths.

10 This strategy would use the explanationist criterion to defend the explanationist criterion. Thus, it may appear circular—but the circularity is not obviously vicious. Rather, it is a case of "rule-circularity"—a form of argument in which the conclusion makes a claim about an inferential rule employed in the argument. Rule-circular arguments, unlike ordinary circular arguments, are plausibly nonvicious (see e.g., Braithwaite 1953).

11 See Nado (forthcoming) for some further preliminary empirical arguments in support of the claim that intuition is heterogeneous.

12 Jackson's full account is considerably more complicated than this, due to his use of two-dimensional semantics. The details arising from the two-dimensional account, however, are not relevant for the purposes of this chapter.

References

Bealer, G. (1992), "The incoherence of empiricism." *Proceedings of the Aristotelian Society Supplementary Volume*, 66, 99–138.

—(2000), "A theory of the a priori." *Pacific Philosophical Quarterly*, 81, 1–30.

Block, N. (1978), "Troubles with functionalism," in W. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology* (Minnesota Studies in the Philosophy of Science, vol. IX). Minneapolis: University of Minnesota Press, pp. 261–326.

Bonjour, L. (1998), *In Defense of Pure Reason*. Cambridge: Cambridge University Press.

Braithwaite, R. (1953), *Scientific Explanation*. Cambridge: Cambridge University Press.

Cappelen, H. (2012), *Philosophy without Intuitions*. Oxford: Oxford University Press.

Churchland, P. M. (1981), "Eliminative materialism and the propositional attitudes." *Journal of Philosophy*, 78, 67–90.

Cummins, R. (1998), "Reflection on reflective equilibrium," in M. DePaul and W. Ramsey (eds), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Lanham, MD: Rowman and Littlefield, pp. 113–27.

Deutsch, M. (2009), "Experimental philosophy and the theory of reference." *Mind and Language*, 24(4), 445–66.

Devitt, M. (2006), "Intuitions," in V. Pin, J. Galparaso, and G. Arrizabalaga (eds), *Ontology Studies Cuadernos de Ontologia: Proceedings of VI International Ontology Congress*. San Sebastian: Universidad del Pais Vasco, pp. 169–76.

Goldman, A. (1987), "Cognitive science and metaphysics." *Journal of Philosophy*, 84, 537–44.

—(2007), "Philosophical intuitions: Their target, their source, and their epistemic status." *Grazer Philosophiche Studien*, 74, 1–26.

Harman, G. (1977), *The Nature of Morality: An Introduction to Ethics*. Oxford: Oxford University Press.

Huebner, B., Bruno, M., and Sarkissian, H. (2010), "What does the nation of China think about phenomenal states?" *Review of Philosophy and Psychology*, 1(2), 225–43.

Jackson, F. (1982), "Epiphenomenal qualia." *Philosophical Quarterly*, 32, 127–36.
—(1998), *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Clarendon Press.
Knobe, J. and Nichols, S. (2007), "An experimental philosophy manifesto," in J. Knobe and S. Nichols (eds), *Experimental Philosophy*. Oxford: Oxford University Press, pp. 3–14.
Knobe, J. and Prinz, J. (2008), "Intuitions about consciousness: Experimental studies." *Phenomenology and the Cognitive Sciences*, 7, 67–83.
Kornblith, H. (1998), "The role of intuition in philosophical inquiry: An account with no unnatural ingredients," in M. DePaul and W. Ramsey (eds), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Lanham, MD: Rowman and Littlefield, pp. 129–42.
—(2002), *Knowledge and its Place in Nature*. Oxford: Oxford University Press.
Machery, E., Mallon, R., Nichols, S., and Stich, S. (2004), "Semantics, cross-cultural style." *Cognition*, 92, B1–B12.
Nado, J. (forthcoming), "Why intuition?" *Philosophy and Phenomenological Research*.
Pust, J. (2000), *Intuitions as Evidence*. New York: Routledge.
—(2001), "Against explanationist skepticism regarding philosophical intuitions." *Philosophical Studies*, 106, 227–58.
Reverberi, C., Shallice, T., D'Agostini, S., Skrap, M., and Bonatti, L. (2009), "Cortical bases of elementary deductive reasoning: Inference, memory, and metadeduction." *Neuropsychologia*, 47, 1107–16.
Sosa, E. (1998), "Minimal intuition," in M. DePaul and W. Ramsey (eds), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Lanham, MD: Rowman and Littlefield, pp. 257–70.
—(2007), "Experimental philosophy and philosophical intuition." *Philosophical Studies*, 132, 99–107.
—(2009), "A defense of the use of intuitions in philosophy," in M. Bishop and D. Murphy (eds), *Stich and His Critics*. Oxford: Blackwell, pp. 101–12.
Stich, S. (2009), "Replies," in M. Bishop and D. Murphy (eds), *Stich and His Critics*. Oxford: Blackwell, pp. 190–252.
Swain, S., Alexander, J., and Weinberg, J. (2008), "The instability of philosophical intuitions: Running hot and cold on truetemp." *Philosophy and Phenomenological Research*, 76, 138–55.
Weatherson, B. (2003), "What good are counterexamples?" *Philosophical Studies*, 115, 1–31.
Weinberg, J. (2007), "How to challenge intuitions empirically without risking skepticism." *Midwest Studies in Philosophy*, 31, 318–43.

Weinberg, J., Gonnerman, C., Crowley, S., Swain, S., and Vandewalker, I. (2012), "Intuitions and calibration." *Essays in Philosophy*, 13, 256–83.

Weinberg, J., Nichols, S., and Stich, S. (2001), "Normativity and epistemic intuitions." *Philosophical Topics*, 29, 429–60.

Williamson, T. (2004), "Philosophical intuitions and skepticism about judgment." *Dialectica*, 58, 109–53.

—(2007), *The Philosophy of Philosophy*. Oxford: Blackwell.

3

Phenomenal Consciousness Disembodied

Wesley Buckwalter and Mark Phelan

1 Introducing some phenomenal bodies

"Suppose we convert the government of China to functionalism, and we convince its officials to realize a human mind for an hour." This is, of course, the beginning of Ned Block's famed "Nation of China" thought experiment (1978, p. 279). In it, Block asks us to imagine that approximately 1 billion people (roughly the population of China at the time) come together to simulate the inner workings of a normal human brain. Each person is given a two-way radio that enables him or her to communicate with others, much like individual neurons in the brain communicate with other neurons. What's more, this complex communication network is hooked up to a remote, artificial body. As the sensory organs of that body are stimulated, the external state of the body is reported on a system of satellites, visible anywhere within China.

According to a set of specified rules, the Chinese respond to these satellite reports by relaying information and commands to one another, and issuing instructions to the body that cause it to execute various behavioral routines. In a sense, the nation of China has now become a "China brain" hooked up to an artificial body (just by radio waves rather than electrochemical impulses). Arguably, such a system satisfies a purely functional description of a mind. Broadly speaking, functionalism defines individual mental states in terms of sensory inputs, behavioral outputs, and relations to other mental states. So according to this view, the various states of the nation of China would amount to the thoughts, feelings, desires, and so on, of normal human beings.

But could a being of such odd construction be said to have a mind, filled with the same kinds of thoughts, feelings, and desires as us? Part of the original

goal of Block's Nation of China thought experiment was to demonstrate that it could not. In fact, Block finds it doubtful that we will say that the China brain has any mental states at all—least of all any "qualitative states, raw feels, or immediate phenomenological qualities" (1978, p. 281). It simply is not made of the right stuff for such states to be possible. This intuition has been taken by some as evidence against the claim that certain theories of functionalism adequately capture how the mind works.

The significance of Block's thought experiment has been debated. But the basic example has remained influential in both psychology and cognitive science.[1] One prominent position that has emerged from this discussion is the view that only certain sorts of entities are capable of having phenomenally conscious mental states such as felt emotions. In addition to being functionally organized in the right sort of way, entities capable of certain mental states (in particular, phenomenal experiences) it is said, must also have the *right sort of biological bodies*. They must be made of the right stuff.

This traditional view held by certain philosophers of mind about how the mind works has also been explored by others as an important principle of folk psychology guiding ordinary mental state ascription. A number of researchers in philosophy, psychology, and cognitive science have defended weaker or stronger versions of a view that we will call the *embodiment hypothesis* (Knobe 2008; Knobe and Prinz 2008; Gray et al. 2011). Generally speaking the embodiment hypothesis states that unified *biological embodiment* is a major psychological factor that cues ordinary attribution of experiences, feelings, emotions, and so on, to other entities. The strongest version of this view is that phenomenal attribution requires biological embodiment. Weaker versions focus on relative levels of attribution, claiming that phenomenal attributions are more likely to be cued as an entity's biological body becomes more salient.[2]

To begin to get a sense of why many have been drawn to the embodiment hypothesis, it may be helpful to consider a specific example of some work on the role of embodiment on mental state ascriptions to different sorts of entities. For instance, in their work studying people's intuitions about group agents, Knobe and Prinz observe:

> It is a striking fact about group agents that we ascribe to them some types of mental states but not others. We might say that Microsoft intends something

> or wants something or believes something . . . but there are other kinds of ascriptions that we would never make to Microsoft. For example, we would never say that Microsoft was feeling depressed. (2008, p. 73)

Knobe and Prinz conduct several studies and find that this is indeed the case. People are very reluctant to ascribe states like *feeling depressed* to the Microsoft Corporation. They go on to explain this striking fact by appealing to two claims. The first claim is that there are important differences in how people ascribe intentional states (like intending or wanting) on the one hand, and states requiring phenomenal consciousness (like feeling sad or depressed) on the other. The second claim is that attributions of these latter kinds of mental states are "sensitive in a special way to information about physical constitution" (Knobe and Prinz 2008, p. 73).[3]

Strictly speaking, the Microsoft Corporation does have a *physical* body. It has a body in the sense that it has a physical presence. It is comprised of factories built of brick and mortar, office buildings, technical laboratories, as well as researchers and employees spread out all across the globe. But the Microsoft Corporation obviously doesn't have a unified body. It is spatially disconnected and includes many disparate kinds of parts. And while it has individual members that are human, the Microsoft Corporation itself clearly lacks a biological body. So while it has a body in some extended sense, it lacks a unified body comprised, among other things, of flesh and blood. According to Knobe and Prinz, we are reluctant to attribute phenomenal states or subjective experiences to the Microsoft Corporation because Microsoft lacks the right kind of unified biological body.[4]

As far as we know, the embodiment hypothesis about folk psychological judgments has not been endorsed by Block or other traditional philosophers of mind directly. Knobe and Prinz (2008) argue for the hypothesis insofar as facts about physical constitution can explain low phenomenal state ascriptions to group agents. Knobe (2008) and Gray et al. (2011) argue for the hypothesis on the grounds that body salience correlates with higher attributions of phenomenal capacities.[5] But leaving aside questions of exactly who has endorsed specific versions of the embodiment view, we note that work by Block, Knobe, and others has made the general embodiment hypothesis a very attractive view that theorists in both psychology and philosophy of mind might be tempted to accept. The general view that unified biological

embodiment is a major psychological factor that cues ordinary attribution of experiences, feelings, emotions, and so on seems to fit some of the data that has been collected across a number of influential studies on mental state ascription. The view would predict that entities with the right kind of biological body are the ones typically thought capable of having phenomenally conscious experiential states. And conversely it holds that entities without the right kind of biological body, such as robots, groups, and ghosts, are typically attributed phenomenal mental states at only very low levels. Furthermore, the embodiment view would be one way of *explaining* the psychological basis of intuitions such as those in Block's original thought experiment, which some philosophers have used as evidence for key philosophical conclusions in the metaphysics of mind.[6]

Nonetheless, we think that philosophers and psychologists should be slow to accept the general embodiment view. Some of the key thought experiments and empirical studies that have been presented to date suggest that embodiment actually does not play that important a role in the way we ordinarily attribute mental states. Specifically, a number of philosophers and cognitive scientists sympathetic to *functional accounts* of the mind have suggested that intuitions favoring embodiment in both thought experiments, like Block's Nation of China, as well as empirical studies on mental state ascription might in fact be trading on subtle cues and distractions related to the *functional organization* of the target entities. On the philosophical side, Dennett (1991), for example, argues that the China brain does not constitute an acceptable counterexample to functionalism. Instead, he argues that Block's thought experiment unfairly relies on a "misdirection of the imagination" because it nonchalantly invites readers to accept that the China brain is complex enough to satisfy the functional roles associated with particular mental states while discouraging them from actually thinking about the complexity of those functional roles. And similarly, on the empirical side, we have argued elsewhere (Phelan and Buckwalter forthcoming), that many of the experimental materials researchers have used to study the influence of embodiment on mental state ascriptions include potential confounds. For example, many include subtle but crucial functional descriptions of entities and their environments (e.g., information about inputs, outputs, and other mental states to which the entities are subject). In addition to considerations about unified biological embodiment,

the inclusion of these potential confounds makes it difficult to assess existing research purported to support the embodiment hypothesis.

Before continuing, it may be helpful to pause and consider how functional information might influence mental state ascriptions. Suppose, for instance, someone was trying to figure out whether or not an entity (let's call this entity "Bob") feels happiness or anger about some state of affairs (such as current low interest rates). In assessing whether Bob is happy or angry about low interest rates, information about Bob's other mental states will be important. Bob is more likely to be happy if he wants to borrow a large sum of money; more likely to be angry if he wants to make money as a lender. Our assessment of Bob's emotional state will also be affected by our beliefs concerning the external stimuli to which Bob is subject. If for instance Bob is angry about the low rates, but then hears a newscast that they've just increased, we will likely temper our assessment of Bob's anger accordingly. Finally, Bob's overt behavior will factor into our assessment of his mental states. If we see Bob cursing or tearing up his lender's agreement, we will be more likely to conclude Bob is angry over the low interest rates. It strikes us as obvious that such functional information contributes to mental state assessment. Indeed, we have shown in prior work that people often consult these kinds of cues when deciding whether or not to ascribe phenomenal states (Buckwalter and Phelan 2013).

Of course, even if functional information of this sort does cue phenomenal state attributions, it could still be that embodiment (for instance, whether "Bob" is a normal human being, a group, or an immaterial ghost) constitutes an important ascription cue as well. However, we think there are good reasons to reject the embodiment hypothesis. This is what we will attempt to demonstrate. In the remainder of this chapter, we examine people's ascriptions of experiential states to entities lacking a biological body. Our goal is to see if ascriptions of phenomenal states to these sorts of entities differ from ascriptions made about normal human beings, or if they tend to work in the same basic way.

For this task, a number of different disembodied entities might have been used. We chose to begin our examination of disembodied ascription with the *phantasmally* disembodied—ghosts and spirits. Lacking in any body whatsoever, spirits constitute the ultimate test of the embodiment view. If the embodiment hypothesis were correct, and embodiment were a crucial cue for phenomenal state attribution, then we would expect important differences in

ascription between human beings, on the one hand, and disembodied ghosts and spirits, on the other—just as we expect to find important differences in phenomenal state attribution for functional information. If functional information—information about the goals, desires, and so forth, of an entity—tends to cue mental state ascription independently of whether the entity has a unified biological body, then it undermines the embodiment hypothesis. This is what we set out to investigate, using spirits as our medium.

2 Previous studies on mental state attributions to disembodied entities

We are under no illusion that our investigation into how people attribute mental states to the disembodied is unprecedented. Several influential studies on God and ghosts have already been conducted. Indeed, previous findings seem to offer contradictory evaluations of the embodiment hypothesis. Some have taken the findings of Gray et al. (2007) to support embodiment.[7] Gray et al. analyzed comparative attributions of a range of mental states to a cast of "characters" ranging from babies to adults, from robots to animals to the dead, and including the ultimate disembodied entity: God. They found that people were less willing to attribute phenomenal mental states (such as feeling fear, hunger, or pain) to God than to many of the other characters in their study. This finding is consistent with the embodiment view, since, presumably, God is thought to lack a body whereas the other characters are not. Furthermore, insofar as lack of embodiment is what explains these low attributions to God, the findings are inconsistent with a simple functionalist account of people's phenomenal state attributions. However, as we have argued elsewhere (Phelan and Buckwalter forthcoming), the findings are readily explicable in functionalist terms, since God is thought to be the ultimate being, who wants for nothing. Surely then he will be thought to suffer fear, hunger, and pain less often than a child or a toad (two other of Gray et al.'s characters).[8]

On the other hand, findings from Jesse Bering and colleagues could be interpreted as challenging the embodiment view. Bering found that adults thought psychological functions—including emotional states—continued after biological death in an agent killed on his daily commute (2002). Bering

and Bjorklund (2004) found a similar pattern for children, who continued to attribute emotional and other mental states to a mouse after it was eaten by an alligator. And Bering et al. (2005) found that both secularly and parochially educated children under ten were proportionally more likely to disagree with statements indicating that psychological functions including emotional states ceased at death. On the assumption that each population thought of the recently deceased as disembodied (a supposition supported by the fact that each population tended to think that biological function ceased at death), these studies present prima facie counterevidence to the embodiment view.

While Bering's work is illustrative, for our present purposes it does not constitute a true test of the embodiment hypothesis. For one thing, it doesn't explicitly compare people's attributions of phenomenal states to the disembodied with their attributions of phenomenal states to normal, embodied humans. Thus it might have missed a tendency to attribute mental states in a way consonant with embodiment. For another thing, it doesn't explicitly manipulate function. Thus it offers no comparison between cues related to embodiment and other salient cues of phenomenal state attribution. Finally, Bering's experimental materials don't explicitly inform participants that the recently deceased agents are disembodied, nor do they ask participants whether they conceive of the dead agents in this way. It thus remains a possibility that experimental participants are not equating death with disembodiment in a way that would shed light on the embodiment hypothesis. Therefore we use this body of research on ordinary beliefs about souls as a point of departure for testing the embodiment view.[9]

3 Disembodying ascription

We present five experiments investigating people's willingness to ascribe emotional states to disembodied ghosts and spirits. The take-home message of this section is that there are certain conditions under which people are perfectly willing to ascribe phenomenal states to these kinds of disembodied, nonbiological entities. Using functional information as a comparison, Experiment 1 demonstrates that ascriptions of emotion to ghosts match those of ordinary human beings, so long as either of those entities satisfies the relevant

functional descriptions. Experiment 2 replicates this effect using experimental probes that lack intentional objects with potentially biasing contextual information (see Section 2.2). Experiments 3–4 display commensurate results for attributions to eternally disembodied spirits. And lastly, using an explicit comparison technique, Experiment 5 provides evidence that in making such ascriptions, participants literally attribute the same phenomenal states to disembodied ghosts and spirits as they attribute to ordinary human beings.

3.1 Experiment 1: Disembodied ghost

3.1.1 Methods

We begin with a between-subjects multifactor experiment designed to test the influence of embodiment and functional cues on ascriptions of phenomenal states.[10] Participants in Experiment 1 (N = 158, 85 female, median age = 27) were presented with the following story about Bob, his ex-wife Melissa, and their son Henry. Roughly half of the participants received a version where the functional information (in this case, the goal Bob attempts to bring about) is to cause his son Henry to hate his mom, while the other half saw a condition in which Bob aimed to make Henry happy:

> **[ANGER/HAPPY]** Bob and Melissa have been married for 15 years. After several months of intense bickering and fighting, they decide to get a divorce. Bob moves out of the house, but still tries to spend time with Henry, their ten-year-old son. He also continues to keep close tabs on his ex-wife Melissa. One day, Bob learns that Melissa has started a new romantic relationship. He hires a private investigator to follow the couple, and take photos of them over a romantic dinner. Bob knows that it will [*cause Henry to hate his mom, Melissa, if he learns that she/make Henry incredibly happy if he learns that his mom, Melissa,*] has started a new, meaningful relationship. Suddenly, Bob gets an idea. If he leaves the pictures in Henry's treehouse in the backyard, Henry is sure to find them when he gets home from school that day. So, Bob jumps in his car and drives to Melissa's house.

After reading one of the versions above, half of the participants saw a conclusion to the story where Bob's biological body is made salient:

> **[EMBODIED]** On the drive over however, Bob is in a car accident. Bob emerges from his car and looks over his body. Everything seems to be

> completely fine—his head, legs and arms. But even though Bob has been in an accident, he won't let that deter him from his earlier goal. He takes the pictures out of his car and walks them over to Melissa's house. He carries them over the back fence and into the treehouse, where Henry is sure to see them.

The remaining participants saw a conclusion to the story where Bob had no biological—let alone physical—body at all:

> [**DISEMBODIED**] On the drive over however, Bob is in a fatal car accident and is killed instantly. Bob emerges from his dead body as a ghost. He now has no form at all—no head, no legs, no arms. Instead, he is something like an invisible force or a spiritual presence. Though he has no limbs with which to touch physical objects, Bob can make objects move without touching them, by floating them through the air. But even though Bob is a ghost, he won't let that deter him from his earlier goal. He causes the pictures to rise out of his car and to float towards Melissa's house. He moves them over the back fence and into the treehouse, where Henry is sure to see them.

All participants were then asked to rate their level of agreement with the following three statements regarding what Bob both felt and believed at the end of the story:

> *Belief.* As Bob moves the pictures into place, he *believes* Henry will find them in the treehouse after school.
>
> *Feel Anger.* As Bob moves the pictures into place, he *feels angry* at Melissa for beginning a new relationship.
>
> *Feel Happiness.* As Bob moves the pictures into place, he *feels happy* for Melissa for beginning a new relationship.

Responses were collected on the same seven-item scale anchored with positive and negative agreement terms designed to measure people's willingness to attribute these intentional states (*Belief*) and experiential states (*Feel Anger* and *Feel Happiness*) to Bob.

3.1.2 Results and discussion

We made three main predictions. A large body of prior empirical work has demonstrated that the behavior displayed by an entity is a crucial factor that cues attributions of intentional states to that entity (Heider and Simmel 1944).

So our first prediction was that given Bob's behaviors in the story, participants would signal high levels of agreement with *Belief* across all conditions in the experiment. Second, we predicted that functional information would have a large impact on phenomenal state attribution, whereby people would signal greater agreement with *Feel Anger* in the ANGER condition, and greater agreement with *Feel Happiness* in the HAPPY condition. And lastly, our third prediction was that embodiment would play little to no role in cuing phenomenal state ascription.

All three predictions were borne out. We found that people strongly agreed that Bob believes Henry will find the pictures in the treehouse after school.[11] Second, there were large effects for function on the way people attributed emotional states to Bob—differences in Bob's nonphenomenal mental states (i.e., his goals) made a big difference in the phenomenal states that were attributed to Bob. And third, emotional state attributions appeared completely unaffected by whether or not Bob was embodied.[12] These results are represented in Figure 3.1.

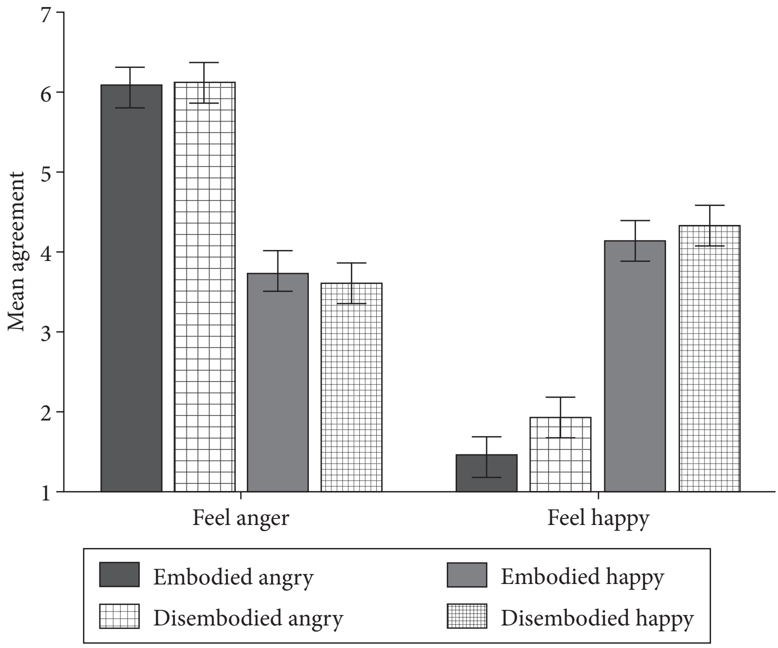


Figure 3.1 Mean agreement with mental state attribution in each condition, grouped by mental state probe. All scales ran 1–7. Error bars +/− SE.

Recall that if the embodiment hypothesis is correct in claiming that a crucial factor that cues phenomenal state attribution is whether or not an entity has a certain kind of biological body, then participants should be more likely to disagree that Bob feels anger or happiness when he exists only as an "invisible force or a spiritual presence" as compared to when his biological body is made salient. But what we saw was that whether or not Bob had or lacked a physical body in the various conditions of the experiment (Figure 3.1: solid vs. patterned bars) seemed to play no role in people's willingness to ascribe these mental states to Bob at the end of the story.

On the other hand, functional cues played a very large role in whether or not people agreed that Bob felt happy or angry. When Bob's goal was to "cause Henry to hate his mom" people were much more likely to agree that Bob feels angry as he places the pictures rather than happy for Melissa for beginning a new relationship. In fact, they attributed this experiential state at roughly the same level as they attributed the intentional state about belief.[13] Lastly, we observe the opposite pattern when Bob wishes to "make Henry incredibly happy." In such conditions, people were much more likely to agree that Bob feels happy rather than angry with Melissa for beginning a new relationship.[14]

3.2 Experiment 2: Disembodied ghost lacking intentional objects

The findings from Experiment 1 begin to motivate the following conclusions. First, when it comes to entities like disembodied souls, having or lacking a human biological body is not utilized as an important cue when attributing phenomenal consciousness. In fact, this information seemed to play no role in people's judgments. Second, participants strongly agreed by comparison that certain entities—embodied or not—can have experiential states when provided with appropriate functional information.

Regarding the first point however, one immediate objection surfaces. Prior work in experimental philosophy of mind has suggested that participants' agreement with phenomenal state attributions are highly sensitive to the amount of contextual information given within experimental probes. Specifically, Arico (2010) found that attributions of phenomenal states to groups that specified an

intentional object (e.g., "Canada's Travel Bureau is experiencing a sudden urge *to pursue internet advertising*") were deemed significantly more acceptable than attributions of phenomenal states to groups that lacked an intentional object (e.g., "Canada's Travel Bureau is experiencing a sudden urge"). In fact, people were much less likely to agree that groups could have a series of phenomenal states when intentional clauses were absent. Thus Arico suggested that the inclusion of an intentional object in experimental probes provides contextual information that can bias phenomenal ascriptions.

Perhaps a similar effect could explain the high ascriptions of phenomenal states to disembodied entities in Experiment 1. It could be that participants attributed emotional states to Bob because the probes that were used included intentional clauses, (e.g., "he feels angry *at Melissa for beginning a new relationship*" vs. "he feels angry"). These clauses might have served to bias disembodied phenomenal state attributions. We conducted our second experiment to rule out this possibility of bias in the probe design.

3.2.1 Methods

Participants in Experiment 2 (N = 147, 53 female, median age = 32) were presented with the same stimulus material combinations as participants in Experiment 1. However, after seeing the vignettes, they were asked to rate their agreement with the following three sentences.[15] These sentences were adjusted to account for the worries above by removing the intentional objects from the probe, thereby limiting the potentially biasing contextual information presented:

Intention. Bob *intends* to move the pictures into place.

Feel Anger No Object. As he moves the pictures into place, Bob *feels angry*.

Feel Happiness No Object. As he moves the pictures into place, Bob *feels happy*.

Responses were collected on the same seven-item scale anchored with positive and negative agreement terms.

3.2.2 Results and discussion

We made two predictions. Our first prediction was that we would replicate each of the results uncovered in Experiment 1. Our second prediction was

that the absence of the intentional phrases and potentially biasing contextual information in the phenomenal state probes in Experiment 2 ("he feels angry" vs. "he feels angry at Melissa for beginning a new relationship") would not result in lower rates of phenomenal state ascription to the disembodied entities in the story.

Both of these predictions were borne out. First, Experiment 2 replicated each effect found in Experiment 1.[16] Participants overwhelmingly attributed *Intention* across the board.[17] Functional cues continued to play a major role in people's judgments. Participants were much more likely to agree with *Feel Anger No Object* in the ANGER condition, and *Feel Happiness No Object* in the HAPPY condition.[18] And lastly, embodiment again seemed to play no role in cuing phenomenal state ascription. Responses in EMBODIMENT and DISEMBODIMENT were nearly indistinguishable. These results can be seen in Figure 3.2.

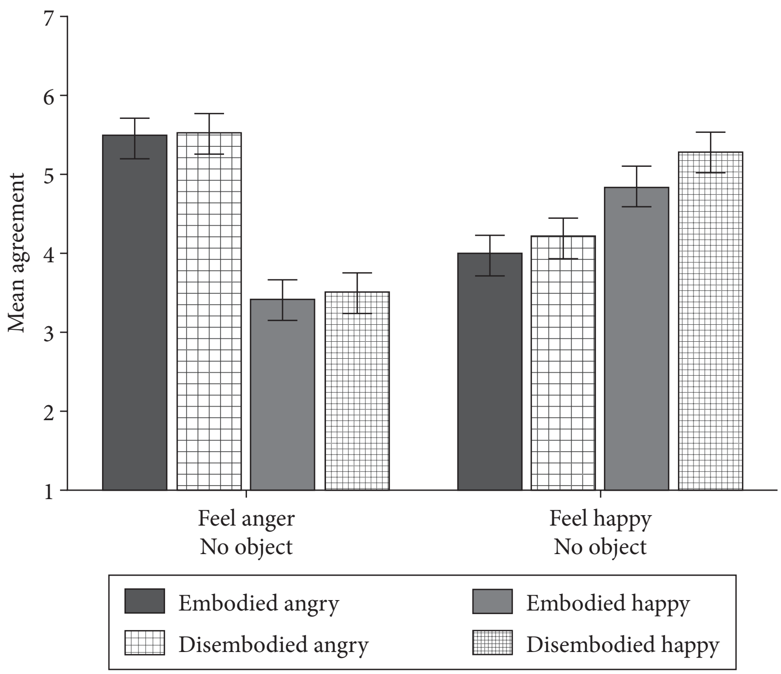


Figure 3.2 Mean agreement with mental state attribution in each condition grouped by mental state probe. All scales ran 1–7. Error bars +/− SE.

Thus Experiment 2 suggests that Arico's results for contextual information bias for group ascriptions do not extend to phenomenal state ascriptions to disembodied entities such as ghosts. In fact, when the intentional object of the phenomenal state probes are dropped in Experiment 2, we again see a clear demonstration of the role that functional information is playing in people's judgments to these entities.

3.3 Experiment 3: Eternally disembodied spirits

Taken together, Experiments 1–2 directly challenge the embodiment hypothesis. But one worry about the entities in these experiments is that participants might be conceiving of them as *nearly* embodied. After all, Bob did not always lack a unified biological body; he was until very recently a normal human being. So perhaps temporal proximity to unified biological embodiment affects people's judgments about the states they attribute to Bob. It could be that there are specific norms related to the genre of ghost stories such that the ghosts of the recently deceased are attributed phenomenal states because they recently possessed human bodies.[19] To rule out these possibilities we conducted Experiment 3 to see if we could replicate the previous results for entities with no temporal proximity to being normal human beings. For this, we turn to spirits that have never been human, or are *eternally* disembodied.

3.3.1 Methods

Experiment 3 mirrored the same between-subjects multifactorial design as Experiments 1–2. Participants (N = 118, 41 female, median age = 30) were presented with cases designed to study the effect of embodiment and functional cues on mental state attribution, this time using an entity that was more purely disembodied. For roughly half of the participants, the story started like this:

> **[EMBODIED]** Fintan is a very private person. He has little connection with the outside world—no computer, no phone, no car. Instead, he hunts or grows his own food with his bare hands. Though he has no money with which to buy tools, Fintan can make many useful objects with the things he finds around him.

The other half saw a story that began as follows:

> **[DISEMBODIED]** Fintan is a nature spirit. He has no form at all—no head, no legs, no arms. Instead, he has always existed as a kind of invisible force or a spiritual presence. Though he has no limbs with which to touch physical objects, Fintan can make objects move without touching them, by floating them through the air.

Both groups then saw the story continue:

> For many years, Fintan has lived in Dirk's Wood beside the Mangahala River. He values the beautiful crystal waters and quiet solitude of the Mangahala above everything else. Recently however, construction has started on the Mangahala Golf Course and Retirement Community. Loggers have begun cutting down segments of Dirk's Wood to accommodate the project, polluting the entire area. Fintan decides that the only way to stop the destruction of his home is to cause their trucks and chainsaws to break in any way he can. And when the loggers bring in more equipment, he breaks that too.

Lastly, to manipulate the functional information specified, participants saw one of two conclusions to the story:

> **[SAD]** But the construction company won't give in. Realizing that there is nothing he can do to stop the loggers, Fintan leaves Dirk's Wood. He must now find another place to call home.
>
> **[HAPPY]** Eventually the construction company gives in. Realizing that he has stopped the loggers, Fintan returns to Dirk's Wood. The place he calls home is now safe.

All participants were then asked the following three questions:

> *Comprehension.* In the story above, Fintan is: [A human being/A nature spirit with no physical body]
>
> *Feel Sadness.* At the end of the story, Fintan feels sad.
>
> *Feel Happiness.* At the end of the story, Fintan feels happy.

Phenomenal state ascription was collected on the same seven-item agreement scale used in Experiments 1–2.

3.3.2 Results and discussion

We made three predictions in Experiment 3. First, we predicted a strong effect for function, whereby people will be much more likely to agree with

Feel Sadness rather than *Feel Happiness* for SAD, and *Feel Happiness* rather than *Feel Sadness* for HAPPY. Secondly, we predicted that embodiment would continue to play no role in people's judgments, even when the object of attribution, Fintan, is a nature spirit that has never occupied a physical body. And thirdly, we predicted that we would conceptually replicate the earlier finding in Experiment 2, that the absence of intentional object clauses does not preclude phenomenal state attribution.

Again all of these predictions were borne out. We found that function made a very large difference to phenomenal state attribution in these cases.[20] People only attributed *Feel Happiness* or *Feel Sadness* in HAPPY and SAD, respectively.[21] Attribution between EMBODIED and DISEMBODIED conditions was indistinguishable. And lastly, these results again persisted despite using phenomenal state probes lacking intentional objects. These findings are displayed in Figure 3.3.

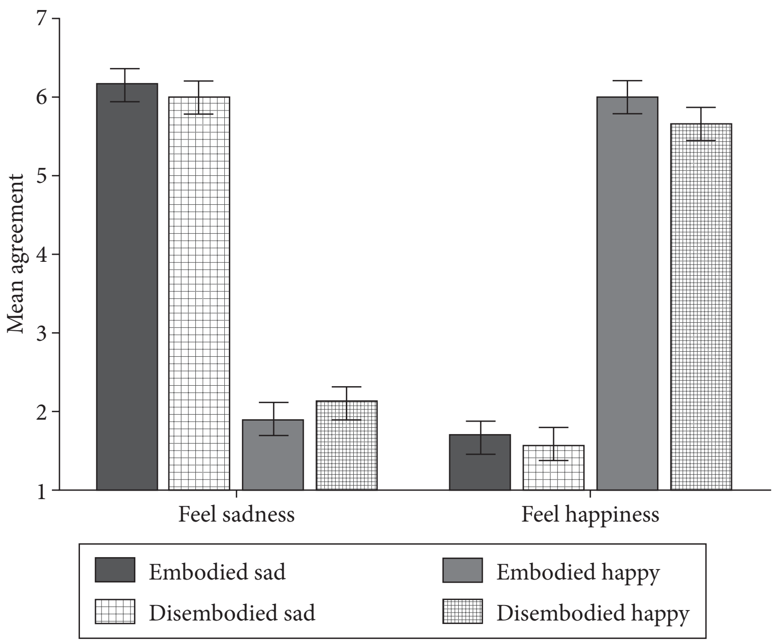


Figure 3.3 Mean agreement with mental state attribution in each condition grouped by mental state probe. All scales ran 1–7. Error bars +/− SE.

3.4 Experiment 4: Eternally disembodied spirits—An alternative measure

One worry about Experiments 1–3 is that they all use the same basic technique for collecting phenomenal state attributions, in which participants were asked two questions about states of opposite valence. But perhaps presenting these two questions together created some undue pressure to ascribe phenomenal states. With this worry in mind, we conducted Experiment 4 using a different measure for state attribution based on confidence judgments.

3.4.1 Methods

Participants in Experiment 4 (N = 120, 37 female, median age = 28) were presented with the same stimulus materials as Experiment 3. However, after seeing the materials, they were asked a different set of questions:

> *Attitude Ascription.* Which do you think best describes Fintan at the end of the story? [Fintan feels sad/Fintan feels happy]
>
> *Attitude Confidence.* How confident are you with the answer you gave to the previous question?

Participants answered *Attitude Ascription* with dichotomous answer choices above. They answered *Attitude Confidence* on a seven-item scale where "1" was anchored with "Not at all Confident" and "7" was anchored with "Extremely Confident." *Attitude Ascription* was then recoded ("−1" for feels sad, and "+1" for feels happy) and multiplied by *Attitude Confidence* to create a combined ascription/confidence score (ranging from −7 to +7) for each of the entities in the various combinations of cases.

3.4.2 Results and discussion

We predicted that this alternative measuring technique in Experiment 4 would still result in the same basic findings as seen in Experiment 3. And that is exactly what we found. Participants still relied on functional information as the crucial cue for ascribing phenomenal states—with total indifference to embodiment.[22] Mean combined scores (Attribution × Confidence Rating) were significantly lower when Fintan fails to save his home, and significantly higher when Fintan succeeds in defeating the loggers.[23] These results are shown in Figure 3.4.

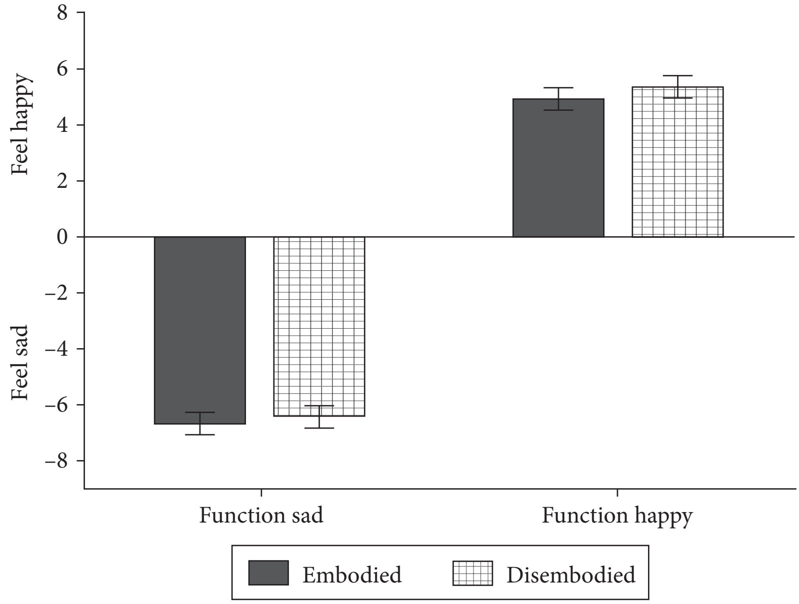


Figure 3.4 Mean combined score (Attitude Ascription × Attitude Confidence) for each type of entity grouped by function. Scores run from −7 to 7. Error bars +/− SE.

3.5 Experiment 5: Spirits, groups, and humans—Explicit emotional comparisons

The previous experiments appear to demonstrate that people often ascribe emotional states to disembodied entities without hesitation—so long as the appropriate functional cues are present. But then again, how can we be sure that participants are applying the phrase "feels sad" to a disembodied entity as they would to a normal human? Perhaps people merely *say* that the spirit feels sad, but *mean* something different than what they mean when they say a human being is sad. In other words, they might not literally attribute the state of sadness to a spirit in the same way they do to a normal human being. In that case, let's say they make an *antirealist* ascription.

To ensure that participants are literally ascribing phenomenal states in both cases, we need evidence that when people ascribe emotional states to ghosts and spirits, they mean to attribute the same emotional states they attribute to other human beings when they make similar ascriptions. In other words, we

need evidence of *realist* ascriptions. In our fifth study, we set out to provide such evidence.

As recent experimental work on quantity implicatures (in addition to other work in experimental pragmatics) demonstrates, it is often very difficult to experimentally uncover what people mean by (or how they interpret) particular sentences.[24] However, our task is at least somewhat less daunting since we do not need to uncover what people ultimately mean when they say a spirit is sad. We simply need to demonstrate that people generally mean the same thing by "sad" when they say, for instance, "a spirit is sad" as they do when they say "a person is sad." There may be numerous ways of examining this question. But one straightforward way is just to ask people to evaluate their mental state ascriptions comparatively. In other words, we could simply ask those who were more or less willing to ascribe emotional states to the spirits how similar the emotional states they meant to attribute were to the emotional states they would attribute to a normal person.

Of course, we would expect some variance in individual responses to this question, so we would need to compare responses to a similar question asked of those who more or less agreed with emotional state ascriptions to the human character in our stories as well. And since we were predicting no difference between people's interpretations of emotion words for the spirit or the human, we would also need some other entity to serve as a control, some entity to which people are willing to ascribe emotional states at the verbal level, but to which they do not really mean to attribute the same emotional states they attribute to normal persons.

For this, we turn again to prior work in the experimental philosophy of mind on group ascriptions. Specifically, Phelan et al. (2013) found that people often offer antirealist phenomenal state ascriptions to group entities (e.g., the Boeing Corporation).[25] Thus group entities seem like the perfect control to use in Experiment 5 when checking for realist ascriptions. Recall the Microsoft Corporation example in Section 1. According to the embodiment hypothesis, people should be thinking about group agents in the same way that they are thinking about disembodied spirits. That is, people should be hesitant to make realist phenomenal ascriptions to both sorts of entities because they lack the right kind of body. So in what follows,

we reexamine participants' judgments by asking for explicit comparisons between different emotional states of humans, disembodied spirits, and group entities.

3.5.1 Methods

Participants (N = 194, 75 female, median age = 26) read vignettes similar to those used in Experiments 3–4. Each vignette began with the introduction of a protagonist that was either a spirit, a human, or a group:

> **[SPIRIT]** Fintan is a nature spirit who strives to protect local forests and rivers. He has no form at all—no head, no legs, no arms. Instead, he has always existed as a kind of invisible force or a spiritual presence. Though he has no limbs with which to touch physical objects, Fintan can make objects move without touching them, by floating them through the air. He uses his spiritual abilities to bring an active approach to nature preservation.
>
> **[HUMAN]** Fintan is an individual who strives to protect local forests and rivers. Through hard work and tireless efforts, Fintan works to protect natural areas from development. Though he has little money with which to support his cause, Fintan exploits his own significant technical skills to bring an active approach to nature preservation.
>
> **[GROUP]** FINTAN is an organization set up to protect local forests and rivers. Through charitable donations and the efforts of group members, FINTAN works to protect natural areas from development. Not only does FINTAN support conservation legislation, it also exploits the technical skills of members to bring an active approach to nature preservation.

All participants then read a short description of the character's struggle against a development project; for the spirit it read as follows (with only necessary changes to the character made across other vignettes):

> For many years, Fintan has worked to protect Dirk's Wood beside the Mangahala River. The spirit values the beautiful crystal waters and quite solitude of the Mangahala above everything else. Recently however, construction has started on the Mangahala Golf Course and Retirement Community. Loggers have begun cutting down segments of Dirk's Wood to accommodate the project, polluting the entire area. After an extended struggle, Fintan decides that the only way to stop the destruction of the woods is to cause the loggers' trucks and chainsaws to break in any way possible. When the loggers bring in more equipment, Fintan breaks those too.

Lastly, participants were presented with one of two possible endings to the story, where Fintan is either successful or unsuccessful at thwarting the loggers' effort:

> **[HAPPY]** Eventually the construction company gives in. Realizing that the loggers have been stopped, Fintan celebrates the preservation of Dirk's Wood.
>
> **[SAD]** But the construction company won't give in. Realizing that nothing can be done to stop the loggers, Fintan gives up on Dirk's Wood.

All participants were then asked the following two questions:

> *Comprehension.* In the story above, Fintan is: [A group/A nature spirit with no physical body/A human being].
>
> *Emotional State Attribution.* At the end of the story, Fintan feels [sad/happy].

Finally, those participants who answered *Emotional State Attribution* with "somewhat agree," "agree," or "strongly agree" proceeded to the additional follow-up question designed to measure realist ascriptions:

> *Comparison.* Consider [Fintan the spirit/Fintan the human being/FINTAN the group (over and above the people that constitute it)]. When you say that this [spirit/man/group] feels [sad/happy], how similar is the feeling of [sadness/happiness] to that of a normal person?

Participants responded to this question by rating their level of agreement on a seven-item scale, running from "Not at all Similar" to "Exactly the Same."

3.5.2 Results and discussion

Given the results of Experiment 1–4, we expected high *Emotional State Attribution* in HAPPY and SAD. And that is exactly what we found. The vast majority of participants (over 80% per condition) signalled at least some agreement with *Emotional State Attribution*.[26]

We now move on to an analysis of *Comparison* for those participants (N = 174) agreeing with *Emotional State Attribution.* If the ascriptions people make to the spirit are realist ascriptions—that is, if people mean the same thing when they attribute emotional states to spirits as they do when they attribute emotional states to people—then we would expect (1) high *Comparison* scores in both HUMAN and SPIRIT (2) with no significant differences

between scores in these conditions. And since prior work has suggested that phenomenal state ascriptions to groups are nonrealist ascriptions, we would expect (3) significantly lower *Comparison* scores in GROUP than in both HUMAN and SPIRIT.

Again, this is exactly what we found. Despite high scores, there was no significant difference in *Comparison* between HUMAN and SPIRIT. Participants indicated that they generally mean the same thing by "feeling happy" or "feeling sad" when directed toward a disembodied spirit or a human being. And consistent with prior findings, we also found that *Comparison* judgments were significantly lower in GROUP than they were for both HUMAN and SPIRIT.[27] Results for *Comparison* are shown in Figure 3.5.

These results suggest that people think of the emotional states they attribute to disembodied entities in the same way they think of the emotional states they attribute to human beings. In other words, this is evidence that they think these states are similar to the emotional states of normal people. And they think these states are dissimilar from the emotional states they attribute to groups. Lastly, recall that the embodiment hypothesis predicted that people

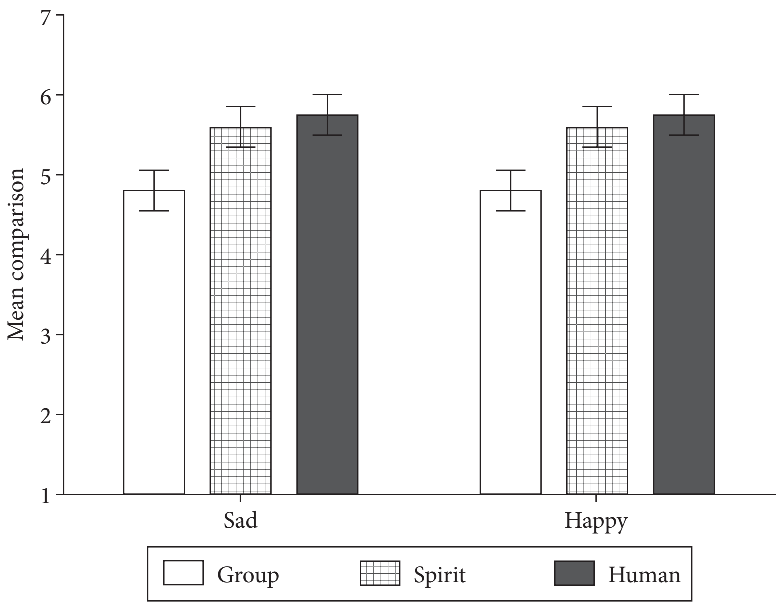


Figure 3.5 Mean comparison judgment for entity type grouped by emotional attribution. All scales ran 1–7. Error bars +/− SE.

would be thinking about groups and disembodied spirits in the same way. But it turns out that we see different results when using our realist measure for comparing ascriptions to these two types of entities.

4 Feeling beyond embodiment

Our experiments suggest that people are perfectly willing to ascribe emotional states to disembodied entities (ghosts and spirits). Though we think more experiments need to be conducted pursuing the question of realist ascription, we think that these results are a promising first step toward the conclusion that findings across Experiments 1–5 constitute strong evidence against the embodiment view. It appears that people really do think that under the right conditions, disembodied entities can have the same kinds of emotional states as human beings. What's more, the data from Study 5 suggest that people think of emotional state ascriptions to disembodied entities *in the same way* they think of emotional state ascriptions to human beings.

Of course, even though participants explicitly state that entities like Fintan are disembodied, it could be that there are specific cultural or social norms which nonetheless suggest that all spirits occupy a location, and thus must possess a body in some indeterminate or minimal sense.[28] Indeed there probably is such a sense in which spirits have bodies, much like there is some mitigated sense in which group entities like Microsoft have bodies. We would only point out that the crucial question—and perhaps the feature that attracted many to the embodiment hypothesis in the first place—was whether or not phenomenal ascriptions are cued in light of possessing a *unified biological body* like our own. It is unclear whether the minimal or indeterminate sense in which ghosts might be assumed to have bodies meets with these criteria.

We should also point out that while we found strong evidence for phenomenal state ascriptions to entities lacking unified biological bodies, embodiment could still have a *relative* impact on ascription. In other words, it's possible that people attribute more, or will be more likely to attribute certain phenomenal states or mental capacities to entities as considerations about the body become *more* salient.[29] While this continues to be a possibility note that in our experiments we found extremely similar rates of ascription between

embodied and disembodied entities. If embodiment made any kind of minimal incremental difference on phenomenal ascription, we did not detect it across our experiments. In contrast, we present strong evidence for the distinct and central role that function plays in ordinary judgments.

These results corroborate previous research by Buckwalter and Phelan (2013) on the important—*otherworldly* even—role that functional information has on the ascription of phenomenal states to diverse sorts of entities. Independently of any influence for body, information about perceptual stimuli, behavioral responses, and other mental states is accompanied by strong attribution of phenomenal states, including, as discussed here, emotional states. One straightforward explanation of this fact is that folk psychology actually identifies phenomenal states with functional roles. This is an interesting question to be pursued in future research on phenomenality and functional role (see Phelan and Buckwalter forthcoming for a discussion of this issue). However, for our current purposes, what we find striking is that people's judgments were highly sensitive to functional role in exactly the same manner for both entities with or without unified biological bodies. We conclude that when it comes to the psychological factors that cue people's actual attributions of phenomenal states to ghosts, perhaps the only apparition here is the embodiment hypothesis itself.

Returning now to the Nation of China thought experiment, it could be that Block's insight about what is ultimately required for phenomenal consciousness is still more or less on the right track. After all, the experiments we conducted only speak to the principles of folk psychology that guide ordinary ascriptions of phenomenal states. They don't rule out the metaphysical possibility that cognition requires some sort of embodiment. While this remains a possibility, we would only note that part of the argument for this metaphysical picture of the mind was motivated by the intuition that the China brain does not have phenomenal states in the first place. But if our results for disembodied entities are shown to be sufficiently general, this intuition may not be widely shared.

Acknowledgments

We are grateful to Matthew Kieran, Joshua Knobe, Aaron Meskin, Shaun Nichols, Justin Sytsma, John Turri, and Joshua Weisberg for helpful comments on earlier versions of this chapter.

Notes

1 Much the same could be said for both the original purpose and subsequent legacy of Searle's "Chinese Room" argument (1980) with respect to the questions of function and embodiment.

2 We use the terms "phenomenal consciousness" and "phenomenal attribution" throughout the chapter when referring to states typically classified by philosophers as qualitative states or states of subjective experience. However, there are some doubts in the experimental philosophy of mind literature concerning whether people have the concept of phenomenal consciousness (see Sytsma and Machery 2010). We emphasize that none of our main arguments or findings here depend on whether or not nonphilosophers draw the phenomenal/nonphenomenal distinction when ascribing experiences, feelings or emotions to disembodied entities, and set the issue aside. We thank Justin Sytsma for discussion on this point.

3 For alternative explanations of Knobe and Prinz's findings, see Phelan et al. (2013), Sytsma and Machery (2009), and our discussion in Section 3.5.

4 Knobe and Prinz seem to focus on the disunity of *corporate* entities as the crucial factor, since one of their later studies suggests that an enchanted chair with a unified "body" can have phenomenal states.

5 This work focuses on the psychological cues for the attribution of mental capacities, while the experiments we present below focus on the cues for attribution of specific mental states. More research is needed to study the subtle differences between these two closely related research questions.

6 After all, the nation of China is one (special kind) of group entity. (Though see Phelan et al. (2013) for an independent source of resistance to thinking that groups really have minds.)

7 Note, however, that Gray and Wegner (2010) question whether these prior findings about God are best interpreted as supporting embodiment.

8 Similar considerations, we think, explain Gray et al.'s findings for other phenomenal states.

9 These are not criticisms of Bering's work, since he wasn't out to investigate the issue of physical realizers at all. In fact, Bering is one of a number of theorists arguing for a particular view about the source of afterlife beliefs, which Bering (2011) calls the *simulation constraint hypothesis*. Nichols (2007), another proponent of the view, encapsulates the basic idea as follows: "part of the reason we believe in immortality is that we can't imagine our own nonexistence" (p. 216). Interesting as the connections are between the embodiment hypothesis and afterlife belief, we set them aside.

10 Participants in this and all subsequent experiments were recruited and tested using commercially available online platforms (Qualtrics and Amazon Mechanical Turk). Participants were located in the United States, and over 85 percent reported English as a first language. They were paid between \$0.30 and \$0.45 for their participation. Participants were prohibited from taking more than one study.

11 *Belief* Disembodied-Anger ($M = 6.39$, $SD = 1.05$), Disembodied-Happy ($M = 6.42$, $SD = 0.81$), Embodied-Anger ($M = 6.54$, $SD = 0.70$), Embodied-Happy ($M = 6.17$, $SD = 1.23$).

12 A 2 (Embodiment) $\times$ 2 (Function) MANOVA was used to compare the influence that body salience and function had on the intentional and experiential states: belief, anger, and happiness. The multivariate result was significant only for function, Pillai's Trace $= 0.48$, $F = 41.36$, $df = (3,137)$, $p < 0.001$. The univariate F tests showed there was a significant difference between attributions of *Feel Anger* $F = 97.79$, $df = (1,142)$, $p = < 0.001$), and *Feel Happiness* $F = 101.16$, $df = (1,142)$, $p = < 0.001$ with respect to function. No differences were detected for *Belief*. No main or interaction effects were detected for *Belief*, *Feel Anger*, or *Feel Happiness* with respect to embodiment. We also conducted an additional alternative analysis. Mann-Whitney U test also detected significant differences in *Feel Angry* within both Embodied $U(70) = 178.00$, $Z = -5.29$, $p < 0.001$, and Disembodied $U(71) = 191.00$, $Z = -5.26$, $p < 0.001$ conditions by function.

13 *Feel Anger* Disembodied-Anger ($M = 6.11$, $SD = 1.19$), Disembodied-Happy ($M = 3.58$, $SD = 1.82$), Embodied-Anger ($M = 6.06$, $SD = 0.91$), Embodied-Happy ($M = 3.75$, $SD = 1.71$).

14 *Feel Happiness* Disembodied-Anger ($M = 1.92$, $SD = 1.65$), Disembodied-Happy ($M = 4.31$, $SD = 1.77$), Embodied-Anger ($M = 1.43$, $SD = 0.70$), Embodied-Happy ($M = 4.14$, $SD = 1.68$).

15 For the sake of uniformity in removing intentional objects, we switched the intentional state tested in Experiment 2 from belief to intent.

16 A 2 (Embodiment) $\times$ 2 (Function) MANOVA was used to compare the influence that embodiment and function had on the intentional and experiential states: intentionality, anger, and happiness. The multivariate result was significant only for function, Pillai's Trace $= 0.308$, $F = 20.94$, $df = (3,141)$, $p < 0.001$. The univariate F tests showed there was a significant difference between attributions of Feel Anger $F = 60.54$, $df = (1,146)$, $p = < 0.001$, and Feel Happiness $F = 14.34$, $df = (1,146)$, $p = < 0.001$ with respect to function.

No differences were detected for *Intention*. And, no main or interaction effects were detected for *Intention*, *Feel Anger No Object*, or *Feel Happiness No Object* with respect to embodiment.

17 *Intention* Disembodied-Anger ($M = 6.50$, $SD = 0.83$), Disembodied-Happy ($M = 6.62$, $SD = 0.49$), Embodied-Anger ($M = 6.46$, $SD = 0.56$), Embodied-Happy ($M = 6.38$, $SD = 0.59$).

18 *Feel Anger No Object* Disembodied-Anger ($M = 5.50$, $SD = 1.33$), Disembodied-Happy ($M = 3.49$, $SD = 1.88$), Embodied-Anger ($M = 5.46$, $SD = 1.44$), Embodied-Happy ($M = 3.41$, $SD = 1.62$). *Feel Happiness No Object* Disembodied-Anger ($M = 4.18$, $SD = 1.72$), Disembodied-Happy ($M = 5.27$, $SD = 1.28$), Embodied-Anger ($M = 3.97$, $SD = 1.76$), Embodied-Happy ($M = 4.84$, $SD = 1.44$).

19 We thank Aaron Meskin for discussion on this point. For more on this point, about the kinds of bodies ghosts might be assumed to have, see discussion in Section 4 below.

20 A 2 (Embodiment) × 2 (Function) MANOVA was used to compare the influence that embodiment and function had on the experiential states anger and happiness. The multivariate result was significant only for function, Pillai's Trace = 0.838, $F = 284.96$, $df = (2,110)$, $p < 0.001$. The univariate F tests showed there was a significant difference between attributions of *Feel Sadness* $F = 374.93$, $df = (1,114)$, $p = < 0.001$, and *Feel Happiness* $F = 542.54$, $df = (1,114)$, $p = < 0.001$ with respect to function. No main or interaction effects were detected for *Feel Sadness* or *Feel Happiness* with respect to embodiment. Three participants were removed for failing *Comprehension*.

21 *Feel Sadness* Disembodied-Sad ($M = 6.00$, $SD = 1.29$), Disembodied-Happy ($M = 2.11$, $SD = 1.40$), Embodied-Sad ($M = 6.14$, $SD = 0.79$), Embodied-Happy ($M = 1.89$, $SD = 0.92$). *Feel Happiness* Disembodied-Sad ($M = 1.57$, $SD = 0.97$), Disembodied-Happy ($M = 5.64$, $SD = 1.19$), Embodied-Sad ($M = 1.66$, $SD = 0.77$), Embodied-Happy ($M = 5.96$, $SD = 0.88$).

22 A 2 (Function) × 2 (Embodiment) between-subjects analysis of variance reveals a main effect for the factor of Function, $F(1, 119) = 839.08$, $p < 0.001$. No other effects were detected.

23 Combined score Disembodied-Sad ($M = -6.41$, $SD = 1.05$), Disembodied-Happy ($M = 5.34$, $SD = 2.47$), Embodied-Sad ($M = -6.60$, $SD = 0.81$), Embodied-Happy ($M = 4.81$, $SD = 3.25$).

24 See Noveck and Reboul (2008) for a useful review.

25 Phelan et al. (2013) argue that phenomenal state ascriptions to groups are often *distributivist*, or that people ascribe states to individual group members rather than to the group as a whole over and above its members.

26 Percent agreement with the emotional state ascription per condition: "Sad Spirit" = 87.9%; "Sad Group" = 81.8%; "Sad Man" = 93.9%; "Happy Spirit" = 93.9%; "Happy Group" = 90.6%; "Happy Man" = 96.7%. Overall, people were somewhat more likely to ascribe happiness ($M = 6.22$) than sadness ($M = 5.79$). A 2 (Emotional state) × 3 (Entity Type) between-subjects analysis of variance reveals a main effect for the factor emotional state, $F(1, 193) = 839.08$, $p < 0.05$. No other effects were detected. We set this result aside.

27 A 2 (Emotional State) × 3 (Entity Type) between-subjects analysis of variance reveals a main effect for the factor Entity Type, $F(2, 173) = 10.29$, $p < 0.001$. No other effects were detected. A Tukey HSD test revealed significant differences for people's interpretations of emotional state attributions between GROUP and both SPIRIT ($p = 0.001$) and HUMAN ($p = 0.001$). However, no significant difference emerged for HUMAN and SPIRIT ($p = 0.897$). Six participants were removed for failing *Comprehension*.

28 We thank Joshua Weisberg for discussion on this point.

29 We thank Shaun Nichols for discussion on this point.

References

Arico, A. (2010), "Folk psychology, consciousness, and context effects." *Review of Philosophy and Psychology*, 1(3), 371–93.

Bering, J. M. (2002), "Intuitive conceptions of dead agents" minds: The natural foundations of afterlife beliefs as phenomenological boundary." *Journal of Cognition and Culture*, 2, 263–308.

—(2011), *The God Instinct: The Psychology of Souls, Destiny, and the Meaning of Life*. London: Nicholas Brealey Publishing.

Bering, J. and Bjorklund, D. (2004), "The natural emergence of afterlife reasoning as a developmental regularity." *Developmental Psychology*, 40, 217–33.

Bering, J., Hernández-Blasi, C., and Bjorklund, D. (2005), "The development of 'afterlife' beliefs in religiously and secularly schooled children." *British Journal of Developmental Psychology*, 23(4), 587–607.

Block, N. (1978), "Troubles with functionalism," in W. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology* (Minnesota Studies in the Philosophy of Science, vol. IX). Minneapolis: University of Minnesota Press, pp. 261–326.

Buckwalter, W. and Phelan, M. (2013), "Function and feeling machines: A defense of the philosophical conception of subjective experience." *Philosophical Studies*, 166(2), 349–61.

Dennett, D. C. (1991), *Consciousness Explained*. Boston: Little, Brown and Company.

Gray, H. M., Gray, K., and Wegner, D. M. (2007), "Dimensions of mind perception." *Science*, 315, 619.

Gray, H. M., Knobe, J., Sheskin, M., Bloom, P., and Barrett, L. B. (2011), "More than a body: Mind perception and the nature of objectification." *Journal of Personality and Social Psychology*, 101(6), 1207–20.

Gray, K. and Wegner, D. M. (2010), "Blaming god for our pain: Human suffering and the divine mind." *Personality and Social Psychology Review*, 14(1), 7–16.

Heider, F. and Simmel M. (1944), "An experimental study of apparent behavior." *American Journal of Psychology*, 57, 243–59.

Knobe, J. (2008), "Can a robot, an insect or God be aware?" *Scientific American Mind*, 19, 68–71.

Knobe, J. and Prinz, J. (2008), "Intuitions about consciousness: Experimental studies." *Phenomenology and the Cognitive Sciences*, 7, 67–83.

Nichols, S. (2007), "Imagination and immortality: Thinking of me." *Synthese*, 159(2), 216–33.

Noveck, I. A. and Reboul, A. (2008), "Experimental pragmatics: A Gricean turn in the study of language." *Trends in Cognitive Science*, 12(11), 425–31.

Phelan, M., Arico, A., and Nichols, S. (2013), "Thinking things and feeling things: On an alleged discontinuity in the folk metaphysics of mind." *Phenomenology and the Cognitive Sciences*, 12(4), 703–25.

Phelan, M. and Buckwalter, W. (forthcoming), "Analytic functionalism and mental state attribution." *Philosophical Topics*.

Searle, J. (1980), "Minds, brains, and programs." *Behavioral and Brain Sciences*, 3(3), 417–57.

Sytsma, J. and Machery, E. (2009), "How to study folk intuitions about phenomenal consciousness." *Philosophical Psychology*, 22(1), 21–35.

—(2010), "Two conceptions of subjective experience." *Philosophical Studies*, 151(2), 299–327.

4

Hallucinating Pain

Kevin Reuter, Dustin Phillips, and Justin Sytsma

The standard interpretation of quantum mechanics and a standard interpretation of the awareness of pain have a common feature: Both postulate the existence of an irresolvable duality. Whereas many physicists claim that all particles exhibit particle and wave properties, many philosophers working on pain argue that our awareness of pain is paradoxical, exhibiting both perceptual and introspective characteristics. In this chapter, we offer a pessimistic take on the putative paradox of pain. Specifically, we attempt to resolve the supposed paradox by undermining the reasons offered for accepting the introspective side of the dualism.

Here is how we will proceed. In Section 1, we lay out the primary reasons that have been offered for thinking that our awareness of pain is paradoxical. We first note the reasons given for adopting a perceptual account of pain (apparent location and nociception), then turn to the reasons given for adopting an introspective account (privacy, subjectivity, and the impossibility of pain hallucinations). In Section 2, we note that previous empirical results cast doubt on the first two reasons given in support of an introspective account of pain. In Section 3, we target the third reason, presenting new evidence concerning lay judgments about the possibility of pain hallucinations. We conclude, in Section 4, that while the current evidence supports adopting a perceptual view of pain, it does not support adopting an introspective view, thereby casting doubt on the purported paradox.

1 The paradox of pain

Many prominent philosophers have found pain to be paradoxical.[1] They argue that our awareness of pains show some features that are associated with outwardly directed perception and other features that are associated with inwardly directed introspection. In recent years, Murat Aydede (2006, 2009) and Christopher Hill (2004, 2006, 2009), in particular, have championed this view.[2] In this section, we briefly detail the primary reasons that have been offered for each side of this apparent dualism, focusing on those given by Aydede and Hill. We begin with the perceptual side, then turn to the introspective side.

1.1 Awareness of pain as perceptual

A perceptual account of pain treats our awareness of pains as being outwardly directed: Pains are taken to be features of body parts. When Joe stubs his toe, for example, the pain is taken to be in his toe; and, in feeling the pain, Joe is aware of a feature of his toe.

Advocates of the paradox of pain have offered two primary reasons for thinking that our awareness of pain is in part perceptual. The first reason is that the apparent locations of pains seem to be in body parts, and that ordinary pain talk supports this claim. The basic assertion is that the phenomenology of feeling pains is such that the pains are located in or on parts of our bodies. It is then held that people, in general, just take the phenomenology at face value, reporting pains to have their apparent spatial locations. When people are asked to state the location of their pains, they don't hesitate to locate them in body parts. They answer, for example, "I have a strong pain in my ankle" or "there is a stinging pain in my shoulder." Advocates of the paradox of pain assert that it is part of the semantics of such expressions that they ascribe pains to nonmental locations. And if we reject the semantics of pain talk, then it turns out that "no one has ever made a true claim about the location or intensity of a pain!" (Hill 2006, p. 89).

The second reason offered for thinking that our awareness of pain is in part perceptual focuses on scientific research into nociception. Brain scientists have produced evidence for the claim that the "structures and processes that explain the awareness of pain are fundamentally akin to the structures and

processes that underlie paradigmatic forms of perceptual awareness" (Hill 2004, p. 342). In other sense modalities, we can tell a causal story that begins with the activation of certain receptors, leading to processing in the relevant areas of the brain, and bringing about awareness of the stimuli. But a similar causal story can also be told for the processing of pain stimuli: Noxious stimuli activate nociceptors, leading to processing in relevant areas of the brain, and bringing about awareness of the stimuli. Given that awareness in other sensory modalities, such as vision and audition, has been plainly thought to be perceptual, the similarity of nociception has been taken as evidence that there is a perceptual aspect to the awareness of pain as well.

1.2 Awareness of pain as introspective

In contrast to perceptual accounts, introspective accounts treat our awareness of pains as being inwardly directed: Pains are not taken to be features of body parts; rather, they are understood in terms of mental states. When Joe stubs his toe, for example, the pain is not taken to be in his toe, but in his mind.

Three primary reasons are offered for thinking that awareness of pain is inwardly directed—the privacy of pains, the subjectivity of pains, and the impossibility of pain hallucinations. Before turning to a discussion of these arguments individually, however, it is worth noting that each can be derived from a common claim in philosophical work on pain—that it is impossible to distinguish the appearance from the reality of pain (see, e.g., Bain 2007; Dretske 2006; Kripke 1980). The reason that is typically given for this claim is not based on extravagant thought experiments or sophisticated arguments; rather, it is simply claimed that the lack of an appearance-reality distinction is part of the common-sense (or "folk") conception of pain. Not only do philosophers claim that their own intuitions favor that the appearance and reality of pain cannot come apart, but they typically maintain that this is how the ordinary person thinks and talks about pain as well. As David Lewis (1980, p. 222) puts it: "Pain is a feeling. Surely that is uncontroversial."[3]

Accepting that pains are feelings, it follows that they are private. In standard cases of perception, however, the objects that are perceived are typically taken to be public. Cakes can be seen, smelled, felt, and (most importantly) tasted by everyone. In contrast, the pain in your stomach that results from having eaten

too many cakes cannot be shared by another person. It is only you who has that pain. Another person might "feel your pain"—that person might empathize with your suffering, perhaps even feeling a pain at the corresponding location in her body—but any pain she feels is another pain, not numerically identical with your own. This suggests that awareness of pain diverges from paradigmatic cases of perception. In contrast, the apparent privacy of pains is in line with cases of introspection, the privacy of one's pain mirroring the privacy of one's thoughts and feelings.

Similarly, if we accept that pains are feelings, then it follows that they are subjective states. To put this another way, to be a pain is to be felt; and, conversely, unfelt "pains" are not pains at all. And, indeed, it is held that the stomach ache produced by eating too many cakes is only a pain if it is felt. Again this contrasts sharply with objects of ordinary perception. It is generally held that a cake can exist without being seen, smelled, or tasted (with apologies to Berkeley).

Finally, if we accept that pains are feelings, then it seems that pain hallucinations must be impossible. If there is no appearance-reality distinction for pains, then the appearance cannot pull apart from the reality and our awareness of pains must be veridical. And, in fact, the philosophical consensus supports the conclusion that pain hallucinations are impossible. For instance, Ned Block asserts that "we do not acknowledge pain hallucinations, [i.e.] cases where it seems that I have a pain but in fact there is no pain" (2006, p. 138). Similarly, Hilary Putnam (1963, p. 218) writes:

> One can have a 'pink elephant hallucination,' but one cannot have a 'pain hallucination,' or an 'absence of pain hallucination,' simply because any situation that a person cannot discriminate from a situation in which he himself has a pain *counts* as a situation in which he has a pain, whereas a situation that a person cannot distinguish from one in which a pink elephant is present does not necessarily *count* as the presence of a pink elephant.

And Saul Kripke (1980, pp. 152–3) suggests the same when he states:

> Pain . . . is not picked out by one of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality. . . . If any phenomenon is picked out in exactly the same way that we pick out pain, then that phenomenon is pain.

In contrast, hallucinations are generally thought to be possible in other sensory modalities. As such, accepting the philosophical common sense, it once again appears that awareness of pain diverges from standard cases of perception.

2 Questioning the paradox

In the previous section, we saw that many hold that pain is paradoxical, finding the awareness of pain to have both perceptual and introspective characteristics. Further, we saw that the reasons given in support of the introspective side of the purported dualism follow from the widely accepted claim that there is no appearance-reality distinction for pains. Despite its widespread acceptance, however, this claim has been challenged in recent years. This challenge puts pressure on the support given for adopting an introspective account of pain, which in turn raises doubts about the purported paradox of pain.

In this section, we detail the recent challenge to the received doctrine, focusing on two articles—Sytsma (2010) and Reuter (2011). Together these articles cast doubt on both the privacy and the subjectivity of pains. In the following section, we build upon this critique, presenting new experimental evidence that casts doubt on the claim that pain hallucinations are impossible. We argue that together these studies raise a significant challenge to the received doctrine and give strong reason to doubt the purported paradox of pain.

2.1 Intuitions about privacy and subjectivity

Among the three characteristics that express the supposed appearance-reality distinction of pain—privacy, subjectivity, and the impossibility of pain hallucinations—Sytsma's (2010) studies raise doubts about the first two: He presents the results of empirical studies that tested the judgments of lay people (people with little to no training in philosophy) about the privacy and subjectivity of pains. Of course, the judgments of lay people might well be mistaken, even about the nature of something as familiar as pain. Recall, however, that the support offered for the received wisdom that there is no appearance-reality distinction for pain is that this is part of the folk conception

of pain. As such, the actual judgments of lay people are directly relevant to assessing the support offered for the claim. And Sytsma's studies suggest that *contra* the philosophical consensus, lay people do not generally conceive of pains as being either private or subjective.

In one set of studies, Sytsma investigated the privacy of pains by asking people to consider cases in which two people share part of their body in common. In one case, he asked participants to consider a pair of conjoined twins. Discussing the putative privacy of pains by considering conjoined twins has the advantage of minimizing epistemological confusion, since the constraint that most of us do not share our body with anybody else means that we have a privileged access to our own pains regardless of whether they are (a) conditions of body parts or (b) mental states. To illustrate, consider the following statement by Eric Schwitzgebel (2010):

> It seems you know your own pains differently and better than you know mine, differently and (perhaps) better than you know about the coffee cup in your hand. If so, perhaps that special 'first-person' privileged knowledge arises through something like introspection.

Perhaps Schwitzgebel is right, perhaps he is not. By merely contemplating people's privileged access to their own pains, however, we cannot deduce whether this is a case of perception of conditions of body parts or introspection of mental states. The case of conjoined twins, though, allows us to dissociate the privileged access we have to our bodies from the privileged access we have to our mental states.

To test whether the commonsense conception of pain really treats pains as being private objects of introspection, Sytsma gave naive participants the following scenario:

> Bobby and Robby are conjoined twins that are joined at the torso. While they are distinct people, each with their own beliefs and desires, they share the lower half of their body. One day while running through a park they forcefully kicked a large rock that, unbeknownst to them, was hidden in the grass. Bobby and Robby both grimaced and shouted out 'Ouch!'[4]

After the scenario, Sytsma asked participants to rate whether the twins felt two different pains or one and the same pain. He found that participants were significantly more likely to answer that they felt one and the same pain.

Further, Sytsma found a similar result in a further study using a case of two people attached to the same hand by a mad scientist. These results suggest that lay people do *not* tend to conceive of pains as being private objects of introspection. Rather, people seem to treat their pains as being "private" in ordinary cases simply because no one else is in a position to perceive their pains.

In another set of studies, Sytsma tested whether the ordinary conception of pains treats them as being subjective mental states, such that pains cannot exist unfelt. For example, Sytsma gave naive participants the following vignette:

> Doctors have observed that sometimes a patient who has been badly injured will get wrapped up in an interesting conversation, an intense movie, or a good book. Afterwards, the person will often report that during that period of time they hadn't been aware of any pain. In such a situation, do you think that the injured person still had the pain and was just not feeling it during that period? Or, do you think that there was no pain during that period?

In opposition to the claim that pains cannot exist unfelt, participants were significantly more likely to answer that the person still had the pain, although he was not feeling it, than that there was no pain.

Together, Sytsma's studies suggest against the claim that the privacy and subjectivity of pains is part of the folk conception, casting doubt on the support offered for the introspective view of pain. At the same time, however, it should be noted that there are some philosophers (Hill 2009; Lycan 2004; Papineau 2007), who agree that there are some counterexamples to the claim that pains are considered to be subjective. Not only do they admit that people can be distracted from little pains, suggesting that they still remain present without us being conscious of them, they also consider it plausible to say that pains can wake people up. Despite these counterexamples, Hill doubts very much that this way of speaking "can be said to represent a dominant strand in our common-sense conception of pain" (2009, p. 171). Further, other cases have seemed to some to support the view that pains are considered to be subjective. For example, Aydede (2006) claims that people do not consider pain to be present in cases where there is tissue damage but painkillers prevent the subject from feeling any pain. Nonetheless, while such cases deserve empirical study

of their own, and while there is much more research to be done in this area, we find that the current evidence raises considerable doubts about whether the ordinary concept of pain supports the view that people do not distinguish the appearance from the reality of pain.

2.2 Further evidence against the introspective account

The putative lack of an appearance-reality distinction regarding pain has not only been questioned by experimental studies investigating people's intuitions, but other empirical data also support the view that the appearance and reality of pain can come apart.

The intensity we attribute to properties like saltiness, loudness, and color has a decisive effect on how confident people are in judging that objects really have this property (Lund 1926). A low degree of confidence will often lead people to give introspective statements (Quinton 1956), making claims about the way things appear to them ("the shirt looks blue") rather than directly about the nonmental objects ("the shirt is blue"). The correlation between low signal intensity and introspection pervades all sense modalities, but has not so far been identified for pain.

In a web-based statistical analysis Reuter (2011) demonstrates that people mostly use the phrase "having pain" when they describe strong pains, but have a preference for the expression "feeling pain" when they describe less intense pains. This analysis suggests that people are confident in ascribing pain to a body part only if the pain is sufficiently strong, and thus that they use expressions of pain in an analogous way to expressions in other sense modalities. These empirical results fuel the following argument:

1. Empirical data shows that the intensity of pain has a decisive effect on whether people assert that they have a pain or feel a pain.
2. "Having pain" and "feeling pain" can be identified as objective statements and introspective statements respectively if their use demonstrates a dependency on the intensity of pain.
3. People's ability to make objective and introspective reports on pain depends on them distinguishing the appearance from the reality of pain.
   From (1), (2), and (3) it follows:
4. People distinguish between the appearance and the reality of pain.

As such, the results of Reuter's study offer further support for the claim that people do in fact draw an appearance-reality distinction for pains.

At the same time, one might argue that the intensity effect revealed in Reuter's study can be explained in another way. For example, one might assert that the intensity effect is a brute fact about the English language, or that people merely imitate the way they express different intensities in the traditional sense modalities. However, Reuter et al. (ms) have been able to reproduce the results in the German language. This data strongly reduces the plausibility of the charge that the intensity effect is merely a linguistic effect. Again, we find that the recently collected data provide evidence against the support offered for the received doctrine that there is no appearance-reality distinction for pain.

3 Studies on pain hallucinations

In Section 1, we noted that three main lines of support are offered for introspective views of pain—privacy, subjectivity, and the impossibility of pain hallucinations. We saw in Section 2, however, that recent empirical findings run counter to the view that people by and large think of pains as being private and subjective. While this evidence suggests against introspective accounts of pain, it does not speak directly to the third line of support—the supposed impossibility of pain hallucinations. In this section, we present new evidence against this view.

3.1 Study 1: The possibility of pain hallucinations, within-participants

To test the received doctrine that it follows from the ordinary concept of pain that pain hallucinations are impossible, in our first study we asked naive participants about the possibility of four types of hallucinations—auditory, pain, visual, and olfactory. Each participant was given the following vignette:

> Jane, Jenny, Sarah, and Susan are all participating in a trial for a new antidepressant being developed by a major drug company. The drug company

suspects that the antidepressant will have some strange side effects. Jane, Jenny, Sarah, and Susan have each been taking the drug twice a day for the past week.

Participants were then asked the four questions below, counterbalanced for order. Participants answered each question by selecting either "yes, it is possible" or "no, it is not possible":

1. After taking the antidepressant this morning, Jenny is walking down the street when all of a sudden it feels like there is a pain in her ankle. Is it possible that Jenny merely hallucinated the pain?
2. After taking the antidepressant this morning, Jane is walking down the street when all of a sudden it sounds like there is a police siren on her left. Is it possible that Jane merely hallucinated the police siren?
3. After taking the antidepressant this morning, Sarah is walking down the street when all of a sudden it looks like there is a butterfly on her right. Is it possible that Sarah merely hallucinated the butterfly?
4. After taking the antidepressant this morning, Susan is walking down the street when all of a sudden it smells like there is vomit in the gutter. Is it possible that Susan merely hallucinated the vomit?

Responses were collected online from 170 native English speakers, 18 years of age or older, with at most minimal training in philosophy.[5]

The results of this study are shown in Figure 4.1 below. Most importantly, we found that 55.9 percent of the participants answered "yes, it is possible"

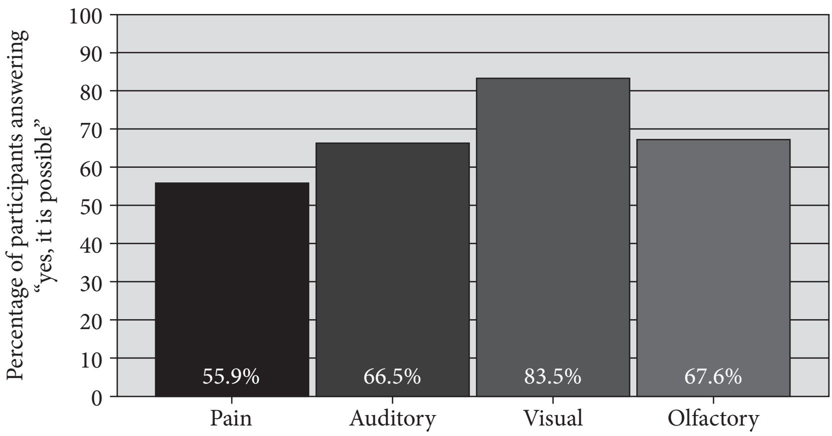


Figure 4.1 Results of Study 1; percentage of participants answering "yes, it is possible" for each of four types of hallucinations.

in response to the pain hallucination question. Thus, while the received doctrine would predict that only a small minority of people would endorse the possibility of pain hallucinations, we found that a majority did so. In fact, this proportion was significantly higher (at the 0.1 level) than the 50.0 percent predicted by chance.[6]

3.2 Study 2: The possibility of pain hallucinations, between-participants

The results of our first study suggest that *contra* the philosophical consensus, people tend to hold that pain hallucinations are possible. In fact, there is reason to believe that, if anything, these results probably understate the case: While there is not generally thought to be any problem with the possibility of auditory hallucinations, only 66.5 percent of the participants in our first study answered "yes, it is possible" for the auditory case. One plausible explanation for this finding is that some participants were hesitant to suggest that the antidepressant might have caused multiple different types of hallucinations. And, in fact, we found that 83.5 percent of the participants answered "yes, it is possible" for the visual case. Thus, it might be that the within-participants design used in our first study served to deflate the numbers for the nonvisual cases.

To test this possibility, we replicated our first study using a between-participants design. In our second study, we gave each participant just one of the four probes below:

> **Pain:** Jenny is participating in a trial for a new antidepressant being developed by a major drug company. The drug company suspects that the antidepressant will have some strange side effects. Jenny has been taking the drug twice a day for the past week.
>
> After taking the antidepressant this morning, Jenny is walking down the street when all of a sudden it feels like there is a pain in her ankle. Is it possible that Jenny merely hallucinated the pain?
>
> **Auditory:** Jane is participating in a trial for a new antidepressant being developed by a major drug company. The drug company suspects that the antidepressant will have some strange side effects. Jane has been taking the drug twice a day for the past week.
>
> After taking the antidepressant this morning, Jane is walking down the street when all of a sudden it sounds like there is a police siren on her left. Is it possible that Jane merely hallucinated the police siren?

Visual: Sarah is participating in a trial for a new antidepressant being developed by a major drug company. The drug company suspects that the antidepressant will have some strange side effects. Sarah has been taking the drug twice a day for the past week.

After taking the antidepressant this morning, Sarah is walking down the street when all of a sudden it looks like there is a butterfly on her right. Is it possible that Sarah merely hallucinated the butterfly?

Olfactory: Susan is participating in a trial for a new antidepressant being developed by a major drug company. The drug company suspects that the antidepressant will have some strange side effects. Susan has been taking the drug twice a day for the past week.

After taking the antidepressant this morning, Susan is walking down the street when all of a sudden it smells like there is vomit in the gutter. Is it possible that Susan merely hallucinated the vomit?

After reading the probe, the participants answered the question by selecting either "yes, it is possible" or "no, it is not possible."

Responses were collected online from 362 participants using the same website and restrictions as in our first study.[7] The results are shown in Figure 4.2. We now found that almost two-thirds of the participants in the pain condition endorsed the possibility of pain hallucinations (64.5%). This percentage is significantly higher than the 50.0 percent predicted by chance.[8] Once again, our results suggest that contrary to what most philosophers claim,

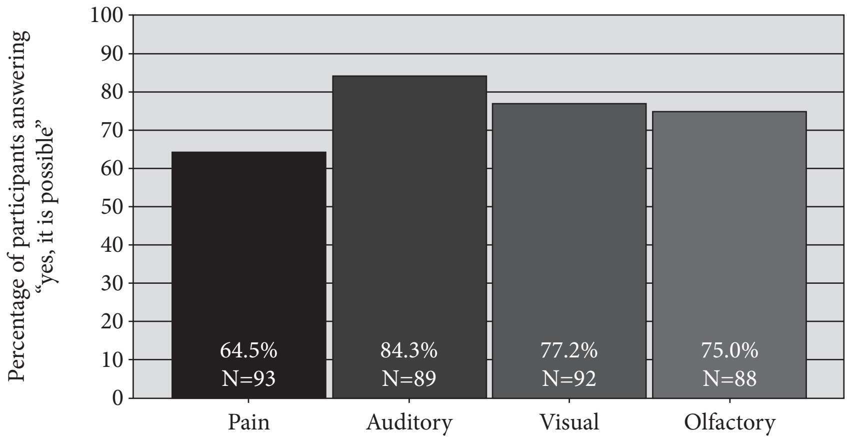


Figure 4.2 Results of Study 2; percentage of participants answering "yes, it is possible" for each of four types of hallucinations.

a significant majority of English speakers believe that pain hallucinations are possible. And this in turn suggests that they hold a concept of pain that allows for an appearance-reality distinction.

At this point, however, it should be noted that the percentage of participants endorsing the possibility of pain hallucinations continues to be lower than the percentage endorsing the other types of hallucinations. Based on this, it might be objected that the lower percentage of positive answers in the pain scenario shows that people are relatively reluctant to endorse the possibility of pain hallucinations. Against this, it should again be noted that participants were in fact *more likely than not* to answer that pain hallucinations are possible. Nonetheless, we do think it is likely that the ordinary concept of pain is not as clear-cut as perceptual concepts for medium-sized dry goods. We deny, however, that this indicates that pains are conceived of as mental states, or that our concept of pain is paradoxical. It seems to us to be more reasonable to take the recent data at face value, acknowledge that a clear majority of people do not believe pains to be private, subjective mental states that cannot be hallucinated, and start to search for new explanations on how to account for the relatively small differences between these perceptual concepts.

One such explanation is the aforementioned constraint that painful body parts are typically connected to only a single mind. It seems likely that this constraint has led to the development of language games that make it more difficult for people to draw an appearance-reality distinction for pains, even though they clearly locate pains in body parts. Another possibility is that pain language reflects the emphasis that we tend to put on the evaluative element in pain judgments. Awareness of pains does not seem to simply involve perception, but also a valence judgment—people find pains to be unpleasant, to one degree or another, and such judgments are reasonably thought to be subjective.[9]

Whereas 83.5 percent of the participants surveyed in our first study affirmed the possibility of having a hallucination in the visual scenario, this dropped slightly to 77.2 percent in our second study. If we assume that the mere possibility of visual hallucinations cannot be seriously challenged, then this figure calls out for explanation. One possibility is that some participants misunderstood the question we asked them. We suspect that some of the discrepancy between the expected and actual result for visual hallucinations

can be accounted for by the ambiguous use of the term "possibility" in everyday talk.[10] For example, when a person asserts that "it is not possible that the mayor will be reelected after the sex scandal," she is not best understood as excluding the theoretical possibility of reelection; rather, she is indicating that she thinks that the event has a low probability of occurring. Applied to our case at hand, we believe that some participants gave negative answers to our questions because they thought that such hallucinations are *unlikely*, not because they thought that they are *impossible*. This means, however, that the real percentage of people believing in the possibility of pain hallucinations will be higher (and not lower) than our result of 64.5 percent indicates because it is reasonable for people to think that a hallucination is possible but unlikely, whereas it does not make sense to consider a hallucination probable but impossible.

3.3 Study 3: Pain hallucinations and pain illusions

In the pain probe used in our second study, we described a case in which it feels to Jenny that she has a pain in her ankle, then raised the question of whether it is possible that Jenny merely hallucinated the pain. Although this question was intended to investigate whether people hold that pains can be hallucinated, some participants might have interpreted the question as asking whether Jenny could have hallucinated the *location* of the pain rather than the pain itself. In other words, instead of answering a question about the possibility of pain hallucinations, some people might have given an answer regarding the possibility of pain illusions. While we doubt that such an interpretation is likely to be widespread (given that we asked them explicitly about the hallucination of the pain and not the localization of the pain), we nonetheless hold that further work is called for here.

The possibility of pain illusions is a more controversial issue in the philosophy of mind than the possibility of pain hallucinations. This is mainly due to the existence of phantom pains, a well-known phenomenon in which people who have had a body part amputated, feel pains that seem to be located in their nonexistent body parts. While some might be inclined to think of phantom limb pains as cases of pain hallucinations, others have argued that they are merely cases of pain illusions and do not undermine the general

claim that there is no appearance-reality distinction for pain. For example, while Hill (2006, p. 76) concedes that phantom pains raise some doubts about whether an appearance-reality distinction is possible for the aspect of location of pains, he downplays the seriousness of the case, insisting that "this is the only discrepancy." He continues: "Thus, while we are prepared to say that the victim's perception that the pain is in the right leg is an illusion, we will allow, and in fact insist, that the pain is in all other respects as it appears to the victim."

Whether or not the perceived location of phantom limb pains is a special case (i.e., a case that does not undermine the more general claim that people cannot distinguish the appearance from the reality of pain) is a question that deserves a full-length paper in itself. We do, however, accept Hill's challenge: If we find that people also consider it possible that other aspects of ordinary pains feel different from the way they really are, then it would seem to be unreasonable to continue to hold that it follows from the ordinary conception of pains that pain hallucinations are impossible.

In order to test Hill's claim, some of us (Reuter et al., ms) conducted a study in which we asked participants the following four questions:

> Q1. Do you think that it is possible to feel a pain as being more intense than it really is?
> Q2. Do you think that it is possible to feel a pain as being less intense than it really is?
> Q3. Do you think that it is possible to feel a pain as being in your ankle even though it is really in your foot?
> Q4. Do you think it is possible to feel a pain as throbbing when it is really burning?

Participants answered by selecting either "yes, it is possible" or "no, it is not possible" for each of the four questions. Responses were collected online from 102 participants using the same website and restrictions as in our previous studies.[11] The results are shown in Figure 4.3. We found that for each question a significant majority of participants answered that the pain illusion is possible.[12] Thus, *contra* Hill, it does not seem that apparent location is the only aspect that draws a wedge between the appearance and the reality of pain on the ordinary conception. In fact, only 5 out of the 102 participants surveyed answered "no, it is not possible" for all four questions.[13]

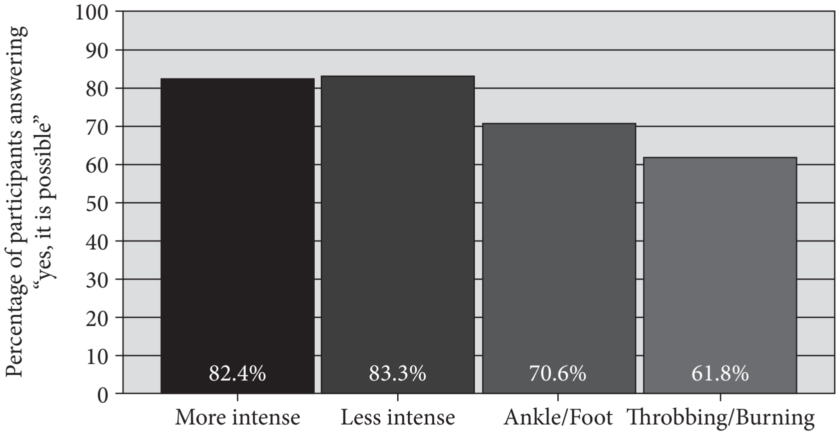


Figure 4.3 Results of Study 3; percentage of participants answering "yes, it is possible" for each of four types of pain illusions.

3.4 Study 4: Understanding "hallucination"

It might be objected that while participants in our first two studies by and large agreed with the claim that it is possible for someone to have a pain hallucination, this might not reflect true agreement: Given the philosophical consensus, it might be thought to be more likely that participants interpreted the term "hallucination" in some other way than we intended. For example, it might be argued that our participants tended to understand talk of pain hallucination in our probe along the lines suggested by Tye (2006), taking pain hallucinations to be hallucinations of tissue damage (which according to Tye's theory is not to be identified with pain itself). Aydede (2009) suggests a similar response when he states:

> Hallucinations or illusions are possible, in one sense, not about feeling/experiencing pain, but about whether these experiences correctly represent some tissue damage, that is, the object of perception in feeling pain.

According to this objection, our participants might have been operating with a similar interpretation when they answered affirmatively to the pain hallucination questions in our first two studies.

It seems to us that this objection already concedes that even in the context of hallucinations, the term "pain" is often interpreted to mean tissue damage, and not the experience of such a bodily state; but, if the term is used to refer

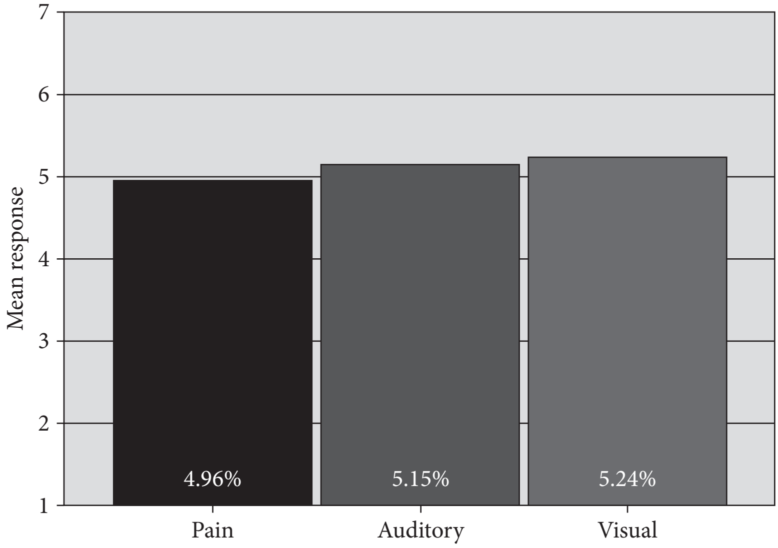


Figure 4.4 Results of Study 4; mean rating for each of three types of hallucinations.

to a bodily state when people localize pain *and* when they think about pain hallucinations, then it seems a fair question to ask: In which situations do people think of pains as being mental states?

This objection can also be tested empirically. To do so, in our fourth study we asked participants to provide a brief description of how they understood the term "hallucination" in addition to asking them whether they agreed or disagreed with each of the three statements below asserting the possibility of a different type of hallucination. The statements were counterbalanced for order. Participants responded by indicating agreement or disagreement with each statement using a seven-point scale anchored at 1 with "Strongly Disagree," at 4 with "Neutral," and at 7 with "Strongly Agree":

> It is possible for someone to have a hallucination of a throbbing pain.
>
> It is possible for someone to have a hallucination of a demonic voice.
>
> It is possible for someone to have a hallucination of a pink elephant.

Responses were collected online from 99 participants using the same website and restrictions as in our previous studies.[14] The results are shown in Figure 4.4. Not only did we find that participants were significantly more likely to agree with the pain hallucination statement than to disagree,[15] but the descriptions

they gave for how they understand the term "hallucination" were in accord with the understanding found in the philosophical literature: A large majority of the participants described hallucinations in terms of a sensory appearance of something that is not really there.[16] The results of this study therefore suggest against the objection. It does not appear that the results of our first study can be explained away in terms of participants having a different understanding of the term "hallucination."

3.5 Study 5: Aydede's challenge

Aydede (2006) raises a more specific semantic challenge to the claim that people generally conceptualize pains as bodily states. He compares the statement "I see a dark discoloration on the back of my hand" with "I feel a jabbing pain in the back of my hand." Aydede argues that while these two sentences have the same surface grammar, they do not have the same truth conditions. He claims that if a person hallucinates the discoloration, then the first sentence is simply false, while the second puts no constraints on the physical condition of his hand.

In order to test Aydede's challenge, we presented participants with a scenario in which a man named John loses one of his hands in a car accident and goes to see a doctor because it sometimes still appears to him that he has a hand. In one case the lost hand visually appears to the person to have a dark discoloration on it; in the other, the lost hand appears to have a sharp pain in it:

> **Dark Discoloration:** John has recently been in a horrible car accident in which he lost his left hand and suffered a severe head trauma. One month later, John honestly reports to his doctor that he often sees his left hand. For example, John told his doctor, 'Right now I see a dark discoloration on the back of my left hand.' But John no longer has a left hand, as the doctor confirms.
>
> When John told his doctor, 'Right now I see a dark discoloration on the back of my left hand,' do you think that his statement was true or false?
>
> **Sharp Pain:** John has recently been in a horrible car accident in which he lost his left hand and suffered a severe head trauma. One month later, John honestly reports to his doctor that he often feels his left hand. For example,

> John told his doctor, 'Right now I feel a sharp pain in the back of my left hand.' But John no longer has a left hand, as the doctor confirms.
>
> When John told his doctor, 'Right now I feel a sharp pain in the back of my left hand,' do you think that his statement was true or false?

Participants were randomly given one of the two scenarios and answered the question by selecting either "TRUE" or "FALSE."

Responses were collected online from 228 participants using the same website and restrictions as in our previous studies.[17] The results are shown in Figure 4.5. What we find is that a significant majority of participants in each condition found the statement to be true—83.3 percent felt that it was true that John felt a sharp pain in the back of his missing hand, while 80.2 percent felt that it was true that he saw a dark discoloration on the back of his missing hand.[18] As such, the results indicate that Aydede is mistaken when he claims that according to the ordinary conception, the statement "I see a dark discoloration on the back of my hand" is simply false in the case of hallucination: A sizable majority of respondents answered that this statement is true, despite John having been described as having lost the hand and, hence, there being no dark discoloration to be seen.

Aydede holds that in the visual case when people realize that they have hallucinated, they correct themselves by switching to talk of the appearance

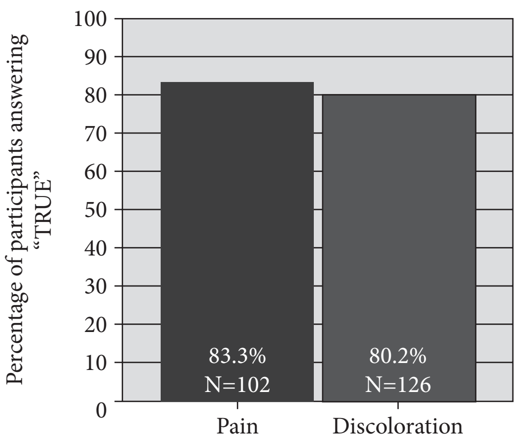


Figure 4.5 Results of Study 5; percentage of participants answering "TRUE" with regard to statements about either a pain or a discoloration in a missing hand.

of a discoloration, but that in the pain case they do not need to make any corrections in their pain reports. The results of our fifth study provide evidence against this view. Further, our results undermine the inference from the premise that people do not correct a statement about feeling a certain pain in a bodily location when they realize they hallucinate, to the conclusion that pains are conceived of as mental states. Why? In the visual case, we do not infer that dark discolorations are mental states (or properties of mental states) even though people take John's statement to be true despite the fact that he is hallucinating. As such, it is at best unclear why we should make a similar inference for the pain case. Put another way, Aydede's challenge depends on the expected difference between responses to the visual case and the pain case; but, there is no such difference (as our data suggests), and thus, his conclusion does not follow.

It is worth noting that the results for the visual case are likely to be quite surprising to many philosophers. We expect that philosophers are likely to think of perceptual verbs like "seeing," "hearing," and "tasting" as success verbs. In fact, Aydede seems to take such a reading of "seeing" for granted in his analysis. He might therefore object to our data and interpretation in two different ways. First, Aydede might claim that the success reading of perceptual words is the only semantically correct reading—people use perceptual terms incorrectly if they violate the success condition. To this objection we would simply respond that Aydede (like most philosophers in this debate) highlights that he is analyzing ordinary concepts. If most people do not use terms like "seeing" in this way, however, then the supposedly "correct" perceptual concepts would not seem to be the ordinary concepts. And, then, Aydede would owe us a new account of why we should think that his understanding of perceptual concepts is correct.

Second, an arguably more promising objection accepts the two alternative readings of the perceptual concept "seeing," but points out that whereas there *is* a common-sense reading of "seeing" that is success-based (even if it is not the only reading), no such reading exists for "feeling pain." This objection, of course, depends on the assumption that despite our results, most people do recognize a success reading for "seeing." We are generally open to this possibility and believe that further study is required to understand when and why people use "seeing" as a success-based concept. The objection,

however, not only claims that there are two possible readings for "seeing," it also states that no success-based reading exists for "feeling pain." Referring back to our data, this objection amounts to saying that those participants (roughly 17%) who respond by saying that the statement "John feels a throbbing pain in his hand" is false are mistaken and make some kind of error in their judgments. We find that the evidence suggests against this assertion.

After presenting the participants in our study with the questions shown above, we also asked them why they responded in the way that they did. Those participants who answered "FALSE" in either of the two scenarios gave remarkably similar explanations of why they believed the presented statement to be false—for example, "he no longer has the limb to feel anything" compared to "[the hand] is not there, so he couldn't see anything." Thus, both sets of responses suggest that there is a success reading not only in the case of "seeing" but also for "feeling pain." This data shifts the burden of proof onto our opponent to explain why we should accept that success-based readings exist for standard cases of perception but not for feeling pain.

4 Conclusion

Many philosophers have found there to be a paradox of pain: They hold that our awareness of pains exhibits both perceptual and introspective characteristics. We are not convinced, however. Specifically, we have doubts about the support offered for the introspective side of the dualism. The support that has been offered primarily rests on claims about the ordinary conception of pain—that it follows from the ordinary conception that pains are private, subjective, and that they cannot be hallucinated. In this chapter we have argued that these claims about the ordinary conception of pain are mistaken. We began by reviewing empirical evidence from Sytsma (2010) and Reuter (2011) suggesting that lay people do not tend to treat pains as being either private or subjective. We then presented the results of five new studies indicating that in contrast to most philosophers, lay people tend to hold that pain hallucinations are possible. Together, these studies provide strong evidence that the ordinary conception of pain is quite different from what philosophers have tended to

claim. And insofar as the case for the paradox of pain depends on claims about the ordinary conception, these studies provide reason to dismiss the purported paradox.

Acknowledgments

This research was funded in part by a Student-Faculty Collaborative Grant from East Tennessee State University. We would like to thank Courtney Oglesby, who was involved in the early stages of this project. We would also like to thank David Harker for his thoughtful comments on an earlier draft of this chapter.

Notes

1 We will follow Hill in talking about "the paradox of pain"; it is worth noting, however, that it might be better (if less elegantly) described as "the paradox of our awareness of pains"—as Hill (2006) and Aydede (2006) have emphasized.

2 These philosophers come to different conclusions, however: Whereas Aydede claims that our awareness of pain is dominated by the introspective strand, Hill favors an eliminativist view on the concept of pain. Other philosophers have also pointed out the dual nature of pain. For example, Michael Tye argues that "the term 'pain' in one usage, applies to the experience; in another, it applies to the quality represented" (2006, p. 101). Similarly, Markus Werning claims: "There are two ways of thinking about pain. [Pain] is itself a state of experience [or] a content of experience." (2010, p. 754).

3 Of course, Lewis is far from alone here. For example, Michael Tye (2006, p. 100) writes that the claim that "pains are necessarily private and necessarily owned is part of our folk conception of pain"; and the obvious explanation offered for this aspect of our common-sense conception "is that pain is a feeling or an experience of a certain sort." Similarly, Aydede (2009) asserts that "the common-sense conception of pain" holds that "pains are sensations with essential privacy, subjectivity, self-intimation, and incorrigibility."

4 In addition, participants were given a second scenario involving a pair of normal undergraduates running a three-legged race for comparison, with the two scenarios being counterbalanced for order.

5 Responses were collected through the Philosophical Personality website (philosophicalpersonality.com). Participants were counted as having more than minimal training in philosophy if they were philosophy majors, had completed a degree with a major in philosophy, or had taken graduate-level courses in philosophy. The participants were 74.7 percent women, with an average age of 41.1 years, and ranging in age from 18 to 84.

6 $\chi^2 = 2.1235$, df = 1, p = 0.07253, one-tailed.

7 The participants were 67.7 percent women, with an average age of 63.6 years, and ranging in age from 18 to 84.

8 $\chi^2 = 7.2688$, df = 1, p = 0.003508, one-tailed.

9 See Sytsma and Machery (2010) and Sytsma (forthcoming) for further discussion of valence with regard to lay mental state ascriptions.

10 People's understanding of the concept of hallucination will be discussed in greater detail below.

11 The participants were 71.6 percent women, with an average age of 36.7 years, and ranging in age from 18 to 75.

12 More Intense: $\chi^2 = 41.4216$, df = 1, $p < 0.001$, one-tailed. Less Intense: $\chi^2 = 44.0098$, df = 1, $p < 0.001$, one-tailed. Ankle/Foot: $\chi^2 = 16.4804$, df = 1, $p < 0.001$, one-tailed. Throbbing/Burning: $\chi^2 = 5.1863$, df = 1, p = 0.01138, one-tailed.

13 It is worth noting that we used the term "feel" in each question in this study rather than the term "appear." We did so because some might find talk of "appearance" to be ambiguous between a phenomenal and a doxastic sense, potentially leading participants to understanding the questions as asking about incorrigibility rather than the possibility of illusions as intended.

14 The participants were 69.4 percent women, with an average age of 44.7 years, and ranging in age from 18 to 82 years.

15 The mean response was significantly above the neutral point: M = 4.96, SD = 2.044, $t(84) = 4.352$, $p < 0.001$, one-tailed. Further, a similar result was found when we removed the question asking participants to describe how they understand the term "hallucination." Responses were collected online from 103 participants using the same website and restrictions as in the previous studies (73.8% women, average age of 43.6 years, ranging in age from 18 to 85 years). Again, the mean response for the pain hallucination statement was significantly above the neutral point of 4: M = 5.03, SD = 1.817, $t(102) = 5.747$, $p < 0.001$, one-tailed.

16 Coding responses using just the key phrases "not there," "aren't really happening," "not real," and "does not exist," 75.2 percent of the participants gave a description in line with that found in the philosophical literature.

17 The participants were 68.4 percent women, with an average age of 40.0 years, and ranging in age from 18 to 70 years.

18 Pain: $\chi^2 = 44.0098$, df = 1, $p < 0.001$, one-tailed. Discoloration: $\chi^2 = 44.6429$, df = 1, $p < 0.001$, one-tailed.

References

Aydede, M. (2006), "Introduction: A critical and quasi-historical essay on theories of pain," in M. Aydede (ed.), *Pain: New Papers on Its Nature and the Methodology of Its Study*. Cambridge, MA: MIT Press, pp. 1–58.

—(2009), "Pain," in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition). http://plato.stanford.edu/archives/spr2013/entries/pain/

Bain, D. (2007), "The location of pains." *Philosophical Papers*, 36(2), 171–205.

Block, N. (2006), "Bodily sensations as an obstacle for representationism," in M. Aydede (ed.), *Pain: New Papers on Its Nature and the Methodology of Its Study*. Cambridge, MA: MIT Press, pp. 137–42.

Dretske, F. (2006), "The epistemology of pain," in M. Aydede (ed.), *Pain: New Papers on Its Nature and the Methodology of Its Study*. Cambridge, MA: MIT Press, pp. 59–74.

Hill, C. (2004), "Ouch! An essay on pain," in R. Gennaro (ed.), *Higher-Order Theories of Consciousness: An Anthology*. Amsterdam: John Benjamins B. V., pp. 339–62.

—(2006), "Ow! The paradox of pain," in M. Aydede (ed.), *Pain: New Papers on Its Nature and the Methodology of Its Study*. Cambridge, MA: MIT Press, pp. 75–98.

—(2009), *Consciousness*. Cambridge: Cambridge University Press.

Kripke, S. (1980), *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lewis, D. (1980), "Mad pain and Martian pain," in N. Block (ed.), *Readings in the Philosophy of Psychology*. Cambridge, MA: Harvard University Press, pp. 216–22.

Lund, F. (1926), "The criteria of confidence." *The American Journal of Psychology*, 37(3), 372–81.

Lycan, W. (2004), "The superiority of HOP to HOT," in R. Gennaro (ed.), *Higher-Order Theories of Consciousness: An Anthology*. Amsterdam: John Benjamins, pp. 93–114.

Papineau, D. (2007), "Phenomenal and perceptual concepts," in T. Alter and S. Walter (eds), *Phenomenal Concepts and Phenomenal Knowledge*. New York: Oxford University Press, pp. 111–44.

Putnam, H. (1963), "Brains and behavior," in R. J. Butler (ed.), *Analytical Philosophy* (2nd Series). Oxford: Basil Blackwell, pp. 211–35.

Quinton, A. M. (1956), "The problem of perception." *Mind*, 64, 28–51.

Reuter, K. (2011), "Distinguishing the appearance from the reality of pain." *Journal of Consciousness Studies*, 18(9–10), 94–109.

Reuter, K., Sytsma, J., and Werning, M. (ms), "Experimentelle Philosphie des Schmerzes."

Schwitzgebel, E. (2010), "Introspection," in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition). http://plato.stanford.edu/archives/spr2013/entries/introspection/

Sytsma, J. (2010), "Dennett's theory of the folk theory of consciousness." *Journal of Consciousness Studies*, 17(3–4), 107–30.

—(forthcoming), "Revisiting the valence account." *Philosophical Topics*.

Sytsma, J. and Machery, E. (2010), "Two conceptions of subjective experience." *Philosophical Studies*, 151(2), 299–327.

Tye, M. (2006), "Another look at representationalism about pain," in M. Aydede (ed.), *Pain: New Papers on Its Nature and the Methodology of Its Study*. Cambridge, MA: MIT Press, pp. 99–120.

Werning, M. (2010), "Descartes discarded? Introspective self-awareness and the problems of transparency and compositionality." *Consciousness and Cognition*, 19(3), 751–61.

5

Taking an "Intentional Stance" on Moral Psychology

Jordan Theriault and Liane Young

A key question of interest to philosophers, and more recently psychologists and neuroscientists, is how we go about attributing complex mental states to the entities (including people) we encounter. Typically, unless we are told explicitly what a person thinks, believes, or desires, our access to information is limited to the observation of behavior. For example, when a man reaches for his pen, we see an arm extended from a body, and fingers wrapped around the pen. Yet we also "see" beyond the surface properties of the action to internal psychological states; we might infer that the man wanted the pen and maybe even for a particular purpose. We might make further behavioral predictions as well (e.g., the man planned to write something down) even in the absence of direct physical evidence (e.g., a piece of paper). By making inferences about mental states that extend beyond what is directly observable, we can be said to have a "theory of mind"—we make inferences about people's internal mental lives (Baron-Cohen 2001; Onishi and Baillargeon 2005; Saxe 2009).

Our focus in this chapter is not to explain theory of mind—how we represent others' specific mental states. Instead we want to focus on the broad categories of how we represent other minds. That is, we want to focus not on what people believe or feel specifically but on whether people have minds that render them capable of believing or feeling. An important means by which we represent other minds is by adopting what Dennett (1987) calls an *intentional stance*. To take an intentional stance is to treat an entity *as though it has a mind*. In taking an intentional stance, we effectively change our approach to dealing with the entity in order to more accurately explain and predict its

behavior. Adopting an intentional stance might be automatic in many cases, driven by low-level cues (e.g., contingent movement, the presence of eyes) that signal goal-directed behavior (Arico et al. 2011; Hamlin et al. 2007; Heider and Simmel 1944; Fiala et al. 2011), but the deployment of an intentional stance can also be consciously motivated, in the absence of external cues, to assist in the prediction of behavior. When an entity behaves in a sufficiently complex manner, the abstraction of an intentional stance allows us to simplify the problem into the following: what would a rational agent do? (Dennett 1987). Importantly, this definition does not require us to apply the intentional stance only to entities that are actually capable of possessing minds. Thus, we can think of a clock or a car as having "a mind of its own" or "its own agenda" even while we understand that certain physical properties might be necessary for mental states to really exist (e.g., a brain).

While an intentional stance suggests that we can treat entities as either having a mind or not, in a dichotomous fashion, recent work in philosophy and psychology has proposed that our representation of minds is not sufficiently captured by one dimension. In fact, our attribution of mental states may be best explained along two dimensions: *Agency* and *Experience* (Gray et al. 2007; Gray et al. 2011; Gray et al. 2012a; Gray et al. 2012b; Gray and Wegner 2011a; Knobe and Prinz 2008; Robbins and Jack 2006; Jack and Robbins 2012). Agency is described as the capacity for purposeful action, and goal-directed behavior (roughly analogous to the target of an intentional stance), while Experience is described as the capacity for sensations and feelings, such as pain and pleasure (Gray et al. 2007; Gray et al. 2011; Gray and Wegner 2011a). For example, we may perceive a robot as capable of forming complex goals, but as less capable of experiencing pain, fear, or other sensations (Gray et al. 2007). By contrast, we might easily attribute sensations to a human baby, even while we recognize its limitations in forming and acting on intentions. These two dimensions of mind attribution—Agency and Experience—may clarify our understanding not just of folk psychology, including our inferences about intentional action, but of folk morality as well (Gray et al. 2012a; Gray et al. 2012b).

In this chapter, we will review the literature on mental state attribution from a psychological perspective. In particular, we will examine the intentional stance (Dennett 1987), where we treat an entity as functionally having a mind, and

expect that it will work toward achieving its goals. We will then explore why an intentional stance alone is insufficient to explain how we attribute minds to others, particularly, why the intentional stance fails to capture our sense of moral concern for others. Moral concern is more completely accounted for by the addition of a phenomenological stance, which, put briefly, is the attribution of emotional experience to others (Robbins and Jack 2006). Paralleling the distinction between an intentional and phenomenological stance are the dimensions of Agency and Experience (Gray et al. 2007; Gray et al. 2011; Gray et al. 2012a; Gray et al. 2012b), which pair moral responsibility with the attribution of Agency, and moral rights with the attribution of Experience. Finally, we will review recent work by Sytsma and Machery (2012), which argues that both Agency and Experience are essential to attributing moral standing (i.e., granting moral rights). This stands in contrast to prior work, which has placed moral standing almost exclusively under the domain of attributions of Experience (Gray et al. 2007; Robbins and Jack 2006).

1 The intentional stance

Taking an intentional stance (Dennett 1987) means predicting and explaining observed behavior in terms of what we know about minds and intentional actors. This approach stands in contrast to cases where we might take a *physical stance* or a *design stance*. In the case of a physical stance, we make predictions based on what we know to be true about physical properties (either through folk physics, or scientific knowledge). Taking a physical stance, if we see one billiard ball roll toward another, we know the first ball will impart force to the second ball and move it. Taking a design stance means making predictions based on the typical functioning of a particular entity, without necessary reference to the entity's underlying physics. If we set an alarm clock to ring in the morning, we predict it will ring at the set time, even without an understanding of the alarm clock's physical operation. The alarm clock has a clear design, acts predictably, and can be dealt with in a way that is independent of its underlying physical properties. Only when our alarm clock breaks down or acts unpredictably do we consider taking it apart to attempt to understand its hardware (Dennett 1981a).[1]

Minds are far too complex for a physical stance or a design stance to make reasonable predictions. Thus, we can adopt an intentional stance, and make predictions based on what we know to be true about minds. Chief among this knowledge is that intentional agents have goals and will work to achieve those goals. Thus, we can expect that agents for whom we adopt an intentional stance will act in whatever way is most likely to achieve their goals. Dennett (1981a) provides the example of a chess-playing computer; to play the computer, we must adopt an intentional stance. Understanding of the physics underlying the operation of the computer—the physical stance—will not be helpful to the player. And to a novice player the means by which the computer selects moves is opaque to the point that he could not predict actions based on his knowledge of how the computer is supposed to work—the design stance. In order to play against the computer, the player adopts an intentional stance, attributing intentions and goals to the computer. The player can then in turn devise strategies that take these intentions into account. Importantly, players adopt an intentional stance despite the fact that the computer lacks the cognitive architecture that we would typically associate with the capacity to represent thought (i.e., a brain). In fact, whether the computer can actually think is irrelevant to the adoption of an intentional stance; from the perspective of the player, treating the computer as though it has a mind is the only means by which he can make predictions about how the computer will behave.

Our utilization of an intentional stance to understand complex behaviors does not necessarily mean that it is appropriate to do so in all cases, or that it will produce the best outcomes. There are certainly cases in which we adopt an intentional stance to our own detriment. The 1997 chess tournament between Garry Kasparov, the world chess champion, and Deep Blue, a chess playing computer designed by IBM (as described by Silver 2012) provides an illustrative example of the potential disadvantages of adopting an intentional stance. In a previous tournament against Deep Blue, Kasparov had consistently taken advantage of his knowledge of how the computer operated: Deep Blue would base its early game strategy on archived data of all previous tournament chess matches that had been played. By playing opening moves that were rarely used in tournaments, Kasparov was able to quickly put the computer in unfamiliar situations. As such Kasparov was making good use of a design stance, where

the actions of the computer could in fact be predicted based on knowledge about how it operates and makes decisions.

Despite this design stance knowledge, Kasparov was allegedly thrown at the end of the first 1997 match, after he had secured an advantageous position and was likely (but not guaranteed) to secure a victory (Silver 2012). Rather than playing to a potential draw, Deep Blue made the unusual and suicidal move of driving a rook into Kasparov's line. Deep Blue forfeited the game after only a few more turns. According to Silver (2012), the strangeness of the move shook Kasparov, as this massive error in a simple position could not be explained by how Kasparov understood the computer to operate. The possible explanations for this bizarre behavior were either that the computer was hiding its capabilities intentionally, that the programmers had thrown the game to make Kasparov overconfident, or else the computer was massively more powerful than previously imagined, and could see ahead so many moves that it had found an alternate route to victory. Either the computer was acting beyond its design as previously understood (by playing mind games), or it was so massively powerful in predicting moves that a simple understanding of its design could no longer suffice (e.g., if it could see more than 20 moves into the future). In the next match, Kasparov famously accused IBM of secretly allowing a grandmaster to make moves on behalf of the machine. Arguably, when the machine failed to conform to the predictions of a design stance, Kasparov adopted an intentional stance toward Deep Blue. In a stroke of irony, the suicidal movement of the rook by Deep Blue was actually a bug: a randomly selected move, which was the result of Deep Blue's indecision. But the damage was done: Kasparov was moved to an intentional stance, and without his knowledge the IBM engineers fixed the bug before the match the next morning (Silver 2012).

That we deploy an intentional stance to predict behavior (or at least attempt it) is further supported by recent work in psychology (Waytz et al. 2010). Waytz and colleagues ran a series of studies in which participants made judgments of the mental capacities of nonhumanoid robots. In their fifth study, the experimenters showed participants a series of videos where robots performed an action, and either asked participants to predict the actions of the robots (with bonus payments for correct answers), or simply asked participants to watch the videos. All participants were shown an initial segment of the video, where the robot performed part of an action (e.g., cleared dishes from the table),

after which the video was paused, and two potential outcomes were presented, (e.g., the robot will either put the dishes in the drawers, or put them on the counter). Participants selected an outcome (or simply read the outcomes in the control condition), and then watched the remainder of the video. Participants then rated the robot on several anthropomorphizing dimensions, including the degree to which the robot had a mind of its own, had consciousness, and possessed intentions, desires, and emotions. Participants who were paid to predict the actions of the robots anthropomorphized them significantly more than the control group. Regardless of whether participants made a conscious decision to take an intentional stance, taking an intentional stance appears to be the consequence of being motivated to understand an entity.

In contrast to Waytz et al. (2010), where an intentional stance was deployed after participants were explicitly instructed to predict behavior, many cases of mental state attribution appear to be driven from bottom-up perceptual features (Arico et al. 2011; Heider and Simmel 1944; Fiala et al. 2011). In other words, the intentional stance is elicited automatically by some external stimuli. Evidence for the bottom-up elicitation of an intentional stance comes from the work of Heider and Simmel (1944), who famously presented "an illusion" in which two smaller shapes were "chased" by a larger one. The shapes were simple geometric figures, sharing few surface features with entities to which we typically attribute mental states. Nonetheless, participants described the short film as depicting a fight between a larger bully triangle and a brave small triangle, a rescue of a small circle by the small triangle, and a furious large triangle smashing up a room in frustration. Describing the shapes in purely mechanical terms would not be incorrect; in fact the patient SM (who suffered from a bilateral amygdala lesion) described the "illusion" in exactly this way, using far fewer affective and social descriptors, and far more movement descriptors than healthy controls (Heberlein and Adolphs 2004). Yet most people cannot help but "see" the shapes as intentional agents—agents who can feel a certain way (e.g., fear, fury) and who can want certain things (e.g., capture, escape).

The largely automatic nature of the attribution of intentional mental states is further supported by its early emergence in development. Hamlin et al. (2007) used shapes (with cartoon eyes) to show 6- to 10-month-old infants a simple story of one shape being helped and hindered respectively by two

other shapes. First, a circle struggles to climb a hill until a square arrives to push him up. Later, the circle struggles to climb the hill again, until a triangle arrives to shove him back down the hill. When infants are later presented with the square and triangle, they prefer to grab the square, presumably due to their understanding of its positive intentions and good character. Dunfield and Kuhlmeier (2010) also demonstrated that by 21 months, infants can understand and draw preferences based on the intentions of adults, even understanding the difference between adults who are accidentally versus intentionally helpful. Infants preferred adults who tried but failed to give a toy to the infant to adults who accidentally provided the desired toy. Thus, our capacity to infer the presence of mental states, and make behavioral predictions based on them, develops early and may reflect a natural source of our social understanding (Waytz et al. 2010).

Adopting an intentional stance—thinking of an entity as having a human-like mind—appears to aid action understanding (or at least perceived action understanding), but this seems to insufficiently capture the full sense of mind we attribute to humans. To humans, we do not only attribute intentions and goals, but also moral rights. Even when we attribute intentions to our chess computer, if we got bored with it then it wouldn't bother us to disassemble it and turn it into a toaster. Mental state attribution might therefore support action understanding on some level as we deal with computers or other inanimate entities, but simply taking an intentional stance or anthropomorphizing an entity does not necessarily imbue it with moral rights. To account for the full extent of our attribution of mind, researchers have begun to converge on the notion that we attribute more than one kind of mind (Gray et al. 2007; Gray et al. 2011; Gray and Wegner 2011a; Jack and Robbins 2012; Knobe and Prinz 2008; Robbins and Jack 2006). In the next section, we will explore how the attribution of moral rights may depend on our ability to empathize and attribute the capacity for pain, pleasure, and emotions.

2 Of two minds: Extending the intentional stance

Philosophers and psychologists have begun to converge on the notion that multiple dimensions of mind attribution compose moral personhood

(Gray et al. 2007; Gray et al. 2011; Gray and Wegner 2011a; Jack and Robbins 2012; Knobe and Prinz 2008; Robbins and Jack 2006). This notion is present even in Dennett's writings (1981b) though not pursued further:

> . . . [When] we declare a man insane we cease treating him as accountable, and we deny him most rights, but still our interactions with him are virtually indistinguishable from normal person interactions unless he is very far gone into madness indeed. In one sense of 'person,' it seems, we continue to treat and view him as a person.

Put plainly, even if we absolve a man of his responsibilities in the case of insanity, we would not excuse ourselves for behaving badly toward him—this person still has rights. This person's pain and pleasure must factor into our moral considerations and how we act toward him.

To account for the observation that moral concern is preserved despite changes in the intentional stance, Robbins and Jack (2006) proposed that we also adopt a *phenomenological stance*. In essence, while the intentional stance involves the attribution of intentions to another person, the phenomenological stance involves the attribution of an emotional experience.

Robbins and Jack (2006) support their dissociation of the intentional and phenomenological stance by drawing attention to a distinction between two sorts of empathy: cognitive empathy and emotional empathy (Davis 1983; Frith 2003; Gonzalez-Liencres et al. 2013; Shamay-Tsoory et al. 2009; Smith 2006). Robbins and Jack (2006) argue that the intentional stance is supported by cognitive empathy, while the phenomenological stance is supported by emotional empathy.

Emotional empathy describes the shared experience of the affective states of others, which is hypothesized to be a phylogenetically older ability (Decety et al. 2012; Gonzalez-Liencres et al. 2013; Smith 2006), with likely evolutionary roots in an ability to empathize with kin and offspring (Trivers 1971). While emotional empathy may have originated in empathy for immediate family, it is hypothesized to have been extended to non-kin over time (cf. expanding the moral circle; Singer 1981). Lesions in brain regions such as the inferior frontal gyrus (IFG) and Brodmann area 44 (which has been implicated as part of the mirror neuron system (Rizzolatti 2005)) have produced deficits in emotional empathy specifically (Shamay-Tsoory et al. 2009). Importantly, emotional

empathy is thought to be distinct from the simple ability to vicariously experience emotional arousal, known as emotional contagion (Lorenz 1935). The distinction could be formalized as follows: emotional contagion takes the form of: "You feel X; therefore I feel X," while emotional empathy takes the form of: "You feel X because Y; therefore I feel X because you feel X" (Gonzalez-Liencres et al. 2013).

Cognitive empathy involves *understanding* the mental states of others, without necessarily *experiencing* the same mental states as the target of the empathy, and could be formalized as follows: "You feel X; therefore I feel Y" (Gonzalez-Liencres et al. 2013). Cognitive empathy is generally considered as theoretically similar to theory of mind, in that both involve an understanding of the representational content of mental states held by others; in addition, measures of cognitive empathy and theory of mind have been shown to correlate (Shamay-Tsoory et al. 2009). Lesions to brain regions that are associated with theory of mind tasks, such as the ventromedial prefrontal cortex (Mitchell et al. 2006), have been associated with deficits in cognitive empathy (Shamay-Tsoory et al. 2009), suggesting that the two processes may also share neural substrates.

The distinction between the intentional stance and the phenomenological stance is supported to some extent by the contrasting deficits in empathy in the cases of autism and psychopathy (Robbins and Jack 2006; Smith 2006). Psychopathy is argued to consist of impaired emotional empathy but preserved cognitive empathy, while autism is argued to consist of impaired cognitive empathy but preserved emotional empathy. Individuals with autism exhibit normal physiological arousal to the distress of others (Blair 1999) and score similar to controls on some emotional empathy tasks (Dziobek et al. 2008); thus, emotional empathy may be preserved. However, the claim that emotional empathy is completely spared in persons with autistic spectrum disorders (ASD) is controversial (Gonzalez-Liencres et al. 2013), due both to some evidence that these persons have difficulty identifying basic emotions in faces (Clark et al. 2008), and to the simple fact that deficits vary so widely in ASD that individual presentations rarely adhere to such a sharp categorical boundary. Nonetheless, several other disorders may fit the profile of impaired cognitive empathy but spared emotional empathy, including frontotemporal lobar degeneration (Rankin et al. 2005), bipolar disorder (Cusi et al. 2010;

Shamay-Tsoory et al. 2009), and borderline personality disorder (Harari et al. 2010; Mier et al. 2013; Minzenberg et al. 2006).

By contrast, in the case of psychopathy, emotional empathy is argued to be impaired, while cognitive empathy remains intact (Robbins and Jack 2006). Psychopaths show impaired autonomic arousal to images depicting distress (Blair et al. 1996) but perform well on tasks measuring theory of mind (Blair et al. 1996); psychopaths even perform well on the "reading the mind in the eyes" task, which requires participants to read off emotions from photographs of eyes (Richell et al. 2003). Other work indicates a similar dissociation in children with psychopathic traits (Jones et al. 2010); these participants show intact theory of mind ability, but reduced concern for other people's feelings in hypothetical scenarios (e.g., "you forgot your friend's birthday and made him feel sad"). This deficit in emotional empathy has also been hypothesized to drive psychopaths' relatively lenient judgments of accidental moral violations (Young et al. 2012). Psychopaths judged accidental harms as morally permissible, presumably due to a failure to be moved emotionally by the harmful outcome, compared to control participants (cf. Cushman 2008). This behavioral profile also stands in contrast to that of ASD participants, who deliver particularly harsh judgments of accidental harms (Moran et al. 2011), due to deficits in encoding agents' innocent mental states.

Above we provided evidence that observers may have both bottom-up and top-down routes available to the adoption of an intentional stance: bottom-up through low-level perceptual cues (Arico et al. 2011; Fiala et al. 2011; Heider and Simmel 1944), and top-down through explicit motivation to understand behavior (Dennett 1981a; Waytz et al. 2010). Bottom-up processes can certainly drive the adoption of a phenomenological stance; emotional contagion (Lorenz 1935) is foundational to emotional empathy (Gonzalez-Liencres et al. 2013). However, can the phenomenological stance be deployed through top-down processes as well? Some recent evidence hints that this may be the case, and that participants can successfully adopt a phenomenological stance when instructed, even to neutral faces that lack any salient emotional content. Participants' success in this task is verified through measurable differences in brain activity in areas known to be associated with emotional empathy (de Greck et al. 2012; Nummenmaa et al. 2008). For example, de Greck et al. (2012) instructed participants to view people's

faces (which were angry or neutral) and either to empathize or to make skin color judgments. These researchers found that empathizing activated bilateral inferior prefrontal cortex, known to be part of the mirror neuron system (Kaplan and Iacoboni 2006). Notably, the inferior prefrontal cortex was preferentially activated when participants actively worked to engage emotional empathy (i.e., adopt a phenomenological stance), compared to when participants made skin color judgments, independent of the emotional content of the face. This result hints at the possibility that emotional empathy can be deployed in a top-down fashion.

What is central to the phenomenological account according to Robbins and Jack (2006) is that witnessing suffering or any aversive phenomenological state should be "primitively morally compelling." In other words, feeling moral concern should flow naturally from recognizing an aversive phenomenological state in others. The work reviewed above on emotional empathy demonstrates that we can easily share and understand the emotions of others, but how are we compelled to care? One potential route is through the "pain matrix," a set of brain regions recruited for both the personal experience of pain, and for witnessing the pain of others (Botvinick et al. 2005; Lamm et al. 2011; Jackson et al. 2006; Singer and Lamm 2009; Singer et al. 2004). According to a recent metaanalysis, the pain matrix includes the anterior medial cingulate cortex (aMCC), posterior anterior cingulate cortex (pACC), and bilateral anterior insula (AI) (Lamm et al. 2011). In one common experimental design (cue-based), participants observe an abstract cue (e.g., a colored light), indicating that either they or a partner will receive a painful electric shock (Lamm et al. 2011). Participants could not see, hear, or touch their partner, and the only indication of the pain their partner would feel was the cue. Regardless of whether the cue indicated the participant or their partner would experience pain, activity in aMCC, pACC, and AI was found in common between the self and other-pain trials (Lamm et al. 2011). Furthermore, neural activity in aMCC, pACC, and AI was also found in common between cue-based, other-pain trials and picture-based studies, in which participants witnessed painful events (such as a car door slamming on someone else's fingers) (Lamm et al. 2011). Cue-based paradigms might model a top-down route to the representation of others' pain, containing no perceptual features typically associated with pain, while picture-based paradigms might model a bottom-up route through their use of painful

imagery. If common neural substrates are involved in the processing of our own pain, and the pain of others, then the morally compelling nature of others' pain may have its roots in the repurposing of neural circuitry that had previously compelled us to be concerned with our own aversive experiences.

If empathy can be driven by bottom-up cues at times, but also can be implemented through top-down processes, then are there any limiting cases where we cannot adopt a phenomenological stance toward an entity? According to work by Knobe and Prinz (2008), adopting a phenomenological stance toward group minds may be one such limiting case. Groups refer to organizations composed of individual group members, such as corporations or political parties, and Knobe and Prinz investigated whether people have the intuition that these groups can be treated as having an analogous mind to an individual. They found that people attribute intentional mental states but not phenomenal mental states toward groups.[2] For example, the statement "ACME Corp. *believes* that by opening 20 new stores they can increase revenue" is reported to sound natural; however, we might be reluctant to endorse the statement "ACME Corp. feels *depressed* because the expansion fails to generate the expected revenue." This use of mental states to describe groups is not simply a metaphorical use; according to participants it appears to be considered literal (Arico et al. 2011). Knobe and Prinz suggest that people may apply intentional mental states broadly to predict behavior, whereas phenomenal mental states might be linked to moral concern, and their attribution might be constrained by additional features. One of these features might be the possession of a physical body with which to empathize.

However, recent work has conflicted with conclusions of Knobe and Prinz (2008). Huebner et al. (2010) present evidence that the difficulty in attributing phenomenal states to groups might instead stem from a Western cultural bias toward individualism. When students in Hong Kong answered questions similar to those used by Knobe and Prinz (2008), they were more likely to attribute phenomenal mental states to groups as entities than American students. This finding is consistent with other work showing the emphasis of Western cultures on individuals' identities and the emphasis of Eastern cultures on collective identity. Further, Knobe and Prinz (2008) have been criticized for making direct comparisons between groups and individuals (Sytsma and Machery 2009). In particular, groups cannot perform many of the

actions typically associated with phenomenal mental states. If a corporation is depressed, for example, it is not capable of crying, insomnia, irritability, and so forth, all of which are behavioral consequences that afford attributions of depression to an entity (Sytsma and Machery 2009).

Understanding that we can adopt a phenomenological stance, in addition to an intentional stance, goes a long way toward making sense of how we attribute mental states. Above we reviewed work that suggests that these two stances are somewhat dissociable (in clinical cases of autism and psychopathy), and that the phenomenological stance likely underlies moral concern. In the next section, we will review a line of psychological work that arrived at a similar conclusion: that we attribute two dimensions of mind, and that these dimensions have unique roles in the domain of morality.

3 Dimensions of mind perception

While we have focused on how people understand other entities by adopting an intentional stance or a phenomenological stance, a distinct line of psychological work has approached mental state attribution using a dimensional approach (Gray et al. 2007; Gray et al. 2011; Gray et al. 2012a; Gray et al. 2012b; Gray and Wegner 2009, 2010, 2011a, 2011b). On this approach, mental state attributions can be made along two dimensions, Agency and Experience, roughly equivalent to the adoption of an intentional stance and a phenomenological stance. In an initial demonstration (Gray et al. 2007), participants judged the relative mental capacities of babies, robots, dead people, adult humans, god, and so on, on the extent that they possess a variety of mental capacities (e.g., capacity for exercising self-control, capacity for feeling pain, etc.). After analyzing the dimensions for factors that could best explain underlying patterns across the entire set of data, two primary components emerged, accounting for 97 percent of the variance in the observed data. First, accounting for 88 percent of the variance in the data was Experience, which included items relating to the experience of hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy. Second, accounting for an additional 8 percent of the variance in the data was Agency, which included items relating to self-control, morality, memory, emotion recognition, planning, communication, and thought.

Some judgments, such as liking the entity, saving the entity from destruction, making the entity happy, or perceiving the entity as having a soul, were correlated with both Agency and Experience. However, of central importance was the observation that Agency was uniquely related to punishing the entity for causing harm, while Experience was related to an aversion toward harming the entity. This work may then be thought to provide evidence of a link between dimensions of mind perception, and moral rights and responsibilities: moral responsibilities are associated with Agency,[3] and moral rights are associated with Experience.

If Agency is associated with moral responsibility and Experience is associated with moral rights then in a typical moral violation involving a perpetrator harming a victim, Agency should be attributed to the perpetrator, who we want to hold responsible for his or her actions, and Experience should be attributed to the victim, whose rights we want to defend. Recent work has made the argument that these associations, combined with the template of a typical moral violation (a perpetrator harming a victim), guide our attribution of mental states (Gray and Wegner 2011a, 2011b; Gray et al. 2012a; Gray et al. 2012b). Based on this template, attributions of Agency and Experience might interact, where attributing more Agency to the perpetrator leads to an increase in Experience attributed to the victim, and vice versa. For example, a harm that is perceived as having been committed intentionally is reported to feel more painful[4] (Gray 2012; Gray and Wegner 2008). Gray et al. (2012a) and Gray et al. (2012b) broadly refer to this phenomenon as *dyadic completion*, where observers will infer a perpetrator in the presence of a suffering victim and a victim in the presence of a harmful perpetrator. Consistent with this account, Gray and Wegner (2010) found a significant negative correlation between religiosity and a "suffering index" (the inverse of a national health index) across American states. States experiencing the most "suffering" also reported the highest belief in god (controlling for education and median income). While Gray and Wegner (2010) did not explicitly test the converse (inferring a victim in the case of victimless immoral behavior), there is no shortage of intuitive examples: such as believing that drug use or homosexuality is necessarily harmful, even in the absence of concrete evidence of harm.

Further work on the dimensional framework of Agency and Experience has highlighted the dissociation of Agency and Experience in subclinical

populations, based on differences in ASD, psychopathic, and schizophrenic characteristics. Gray et al. (2011) replicated the results of Gray et al. (2007) and in addition collected measures of individual differences on the Autism Spectrum-Quotient Scale (Baron-Cohen et al. 2001), measuring traits related to ASD; the Self-Report Psychopathy Scale (Paulhus et al. 2009), measuring traits related to psychopathy; and the Schizotypy Personality Questionnaire (Raine and Benishay 1995), measuring traits related to schizophrenia. Participants scoring high on the Autism Scale perceived less Agency in entities; participants scoring high on the Psychopathy Scale perceived less Experience in entities; and participants scoring high on the Schizotypy Personality Scale perceived more Agency in entities. These results have been taken as further support for Agency and Experience as largely orthogonal and dissociable dimensions of mind attribution. These results also parallel the deficits discussed above of cognitive empathy in ASD, and emotional empathy in psychopathy.

4 Agency and experience as sources of moral standing

The work we have reviewed so far suggests that Experience is essential to granting moral rights (Gray et al. 2007; Jack and Robbins 2006), while Agency is essential to attributing moral responsibility (Gray et al. 2007). However, Sytsma and Machery (2012) have recently suggested that Experience alone cannot completely account for the range of ways in which we attribute moral standing (moral rights). Experience works well to account for the sources of moral standing considered by utilitarian thinkers (Bentham 1781/2000; Singer 1981), but on deontological grounds the moral standing of an entity often depends on its rationality (Kant 1785/2005). In other words, Agency may contribute more to moral standing than previously thought. Utilitarianism maintains that moral decisions should be based on a metric, where the correct moral action is the one that maximizes well-being and minimizes suffering (Bentham 1781/2000). This utilitarian metric relies on victims' capacity for suffering and on observers' capacity to both empathize with others, and desire to prevent their suffering. Kant's deontological morality (1785/2005), on the other hand, emphasizes that the moral standing of victims should be based on the extent to which they are capable of making rational decisions, rather than

simply being driven by passions (emotion) (see Sytsma and Machery (2012) for a thorough review of thinkers in utilitarian and deontological schools of thought). For Kant, the source of moral standing appears to be Agency, rather than Experience. On this basis, Sytsma and Machery (2012) argue that moral standing may derive from two sources: Agency and Experience.

Sytsma and Machery (2012) provide an illustrative example of Agency's role in moral standing in the 1550–51 debate in Valadolid, Spain, over whether the Spanish could rightfully enslave the aboriginal Indians in North America. The debate was between Sepúlveda, who argued for the enslavement of the Indians, and Las Casas, who argued against it. Importantly, the debate centered not on the capacity of the Indians to suffer, but on whether the concept of barbarians could appropriately be applied to them. Sepúlveda argued that the Indians were uncivilized; while Las Casas emphasized that the Indians had sophisticated civilizations and languages, and applied their own rule of law (Sytsma and Machery 2012). Las Casas was successful, and a papal decree was issued, declaring that the Indians were not to be enslaved:

> The enemy of the human race, who opposes all good deeds in order to bring men to destruction, beholding and envying (the spreading of the Catholic Faith), invented a means never before heard of, by which he might hinder the preaching of God's word of Salvation to the people: he inspired his satellites who, to please him, have not hesitated to publish abroad that the Indians of the West and the South, and other people of whom We have recent knowledge should be treated as dumb brutes created for our service, pretending that they are incapable of receiving the Catholic Faith.
>
> We, who, though unworthy, exercise on earth the power of our Lord and seek with all our might to bring those sheep of His flock who are outside into the fold committed to our charge, consider, however, that *the Indians are truly men and that they are not only capable of understanding the Catholic Faith but, according to our information, they desire exceedingly to receive it*. (Emphasis added)[5]

The moral status of the Indians in this case did not hinge on whether or not they were capable of suffering, but instead on their rationality. The papacy is broadening the community of the catholic congregation to include the Indians as a result of their capacity and (alleged) desire to share the same set of beliefs.

Experimental evidence provided by Sytsma and Machery (2012) provides further support for the role of Agency in moral standing. Sytsma and Machery (2012) first replicated the previous finding of Experience, but not Agency, driving moral standing (Gray et al. 2007; Gray et al. 2011), using the example of lethal experimentation on monkeys. However, they repeated their initial paradigm in another scenario: would it be acceptable for humans to experiment on the population of a newly discovered alien race? In this case, participants were more opposed to experimentation on alien races with Agency, but the race's capacity for Experience was irrelevant to their moral standing. A follow-up study asked participants about the acceptability of lethal experimentation on a single alien where, in addition to replicating the effect of Agency, a small effect of Experience emerged as well. The results suggest that Experience may play a more prominent role in the moral standing of individuals, rather than the moral standing of groups (cf. Knobe and Prinz 2008).

The potential role of Agency as an additional source of moral standing is promising for interpreting the broad range of moral norms and behaviors. In particular, how moral standing derives from Agency in contrast to Experience could be a particularly fruitful avenue of future research. Witnessing suffering should be "primitively morally compelling" (Robbins and Jack 2006), and we discussed above how this might be so by reviewing the literature surrounding emotional empathy and the vicarious experience of aversive states through the pain matrix. But the mechanism through which Agency grants moral standing remains unclear, and the work reviewed in this chapter seems to suggest that simply attributing Agency is not enough to compel us to grant moral rights to the target (Gray et al. 2007; Gray et al. 2011; Gray et al. 2012a; Gray et al. 2012b). One alternative might be that Agency compels us to take the goals of rational entities seriously (see Kant 1785/2005). If someone were to complain they were being mistreated, and we saw them as rational, then we should either address their complaint, or find a justifiable reason for why we shouldn't have to. In contrast, if a child says being forced to eat broccoli is mistreatment then we should feel comfortable overruling them, as we don't see them as rational to the same extent that we are. Moral standing might then derive from Agency through the reasonable assumption that if an entity were rational, it would want to be treated ethically. For instance, when the Spaniards determined that the Indians were rational, it was not

empathy that was responsible for the Spaniards' restraint, but respect for the expressed desire of the Indians *to not be enslaved.* A counterexample from Douglas Adam's *Restaurant at the End of the Universe* makes the role of the expressed goals of a rational entity even more apparent by providing an Agentic creature that advocates *against* preserving its life. While traveling through space, Arthur Dent, a human, and his fellow travelers arrive at the restaurant at the end of the universe, where they meet the main course: a cow that wants to be eaten (Adams 1980, Chapter 17). Arthur is disgusted by the cow's recommendation of which body parts to consume, and asks to have a green salad.

> 'A green salad?' said the animal, rolling his eyes disapprovingly at Arthur.
>
> 'Are you going to tell me,' said Arthur, 'that I shouldn't have a green salad?'
>
> 'Well,' said the animal, 'I know many vegetables that are very clear on that point. Which is why it was eventually decided to cut through the whole tangled problem and breed an animal that actually wanted to be eaten and was capable of saying so clearly and distinctly. And here I am.'

Clearly the cow's invitation doesn't solve the dietary dilemma for Arthur and indeed introduces a new dilemma between Arthur's aversion to killing the cow (the cow's Experience), and the cow's insistence on being killed and eaten (the cow's Agency). This example reveals that the cow's Agency can influence its moral standing, but factors such as Experience will continue to have influence, even against the cow's expressed desire to be killed.

5 Conclusion

In this chapter we've reviewed the theory behind the intentional stance, and shown that it is central to how we make predictions about complex behaviors. Despite its importance, the intentional stance (or Agency) in isolation cannot completely account for our attribution of moral standing toward entities. Recognizing that we also adopt a phenomenological stance (attributing Experience) toward entities, and that this happens largely independently of our attribution of Agency, provides a more complete picture of our understanding

of mental states. Ultimately, however, moral standing may not be the exclusive domain of either Agency or Experience (Sytsma and Machery 2012), and by recognizing the multiple sources of our moral standing we might come closer to capturing the entirety of our moral universe.

Notes

1 Although actually solving the problem requires us to adopt a *physical* stance toward the clock, Waytz et al. (2010) found that people are even more likely to adopt an *intentional* stance in anthropomorphizing the malfunctioning clock. Thus, when an entity appears to be beyond easy understanding, we may resort to attributing mental states to that entity (e.g., my clock *forgot* to wake me up because it is *out to get me*).

2 More specifically, Knobe and Prinz tested folk intuitions about phenomenal consciousness: the second-order property that there "is something to be like" in states such as seeing red, hearing a C# musical note, etc. (Sytsma and Machery 2010). In particular, Knobe and Prinz wanted to test whether nonphilosophers understand this philosophical concept. Sytsma and Machery (2010) have argued that Knobe and Prinz conflate the folk understanding of this concept with the folk understanding of subjective experience (i.e., the phenomenal stance). We leave aside this debate surrounding phenomenal consciousness and simply present the work by Knobe and Prinz as an illustration of the relative independence with which the intentional and phenomenological mental states can be attributed.

3 One might notice that Agency involves the attribution of moral responsibility, while adopting an intentional stance does not necessarily implicate the target as morally responsible. The boundaries of where the capacity for intentional action gives way to moral responsibility have, to our knowledge, not been well explored. However, based on the history of moral philosophy exploring the moral consequences of being a rational, self-determining actor (Sytsma and Machery 2012), we suggest that the boundary has to do with seeing a target as capable of rationally forming its own goals. Remember that adopting an intentional stance allows us to assume a target will act rationally to achieve its goals, but that does not necessarily mean it can choose them. For example, the chess computer certainly never chose to desire to win at chess.

4 The claim that the Agency of the perpetrator affects the subjective experience of pain in the victim is beyond the focus of this chapter. To reiterate, we are specifically concerned with the attribution of mental states to others as opposed to the subjective experience of one's own mental states.

5 http://www.papalencyclicals.net/Paul03/p3subli.htm

References

Adams, D. (1980/2010), *The Restaurant at the End of the Universe*. London: Pan Macmillan Limited.

Arico, A., Fiala, B., Goldberg, R. F., and Nichols, S. (2011), "The folk psychology of consciousness." *Mind & Language*, 26(3), 327–52.

Baron-Cohen, S. (2001), "Theory of mind in normal development and autism." *Prisme*, 34, 174–83.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001), "The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians." *Journal of Autism and Developmental Disorders*, 31(1), 5–17.

Bentham, J. (1781/2000), *An Introduction to the Principles of Morals and Legislation*. Kitchener: Batoche Books.

Blair, R. J. R. (1999), "Psychophysiological responsiveness to the distress of others in children with autism." *Personality and Individual Differences*, 26, 485.

Blair, R. J. R., Sellars, C., Strickland, I., Clark, F., Williams, A., Smith, M., and Jones, L. (1996), "Theory of mind in the psychopath." *Journal of Forensic Psychiatry*, 7, 15–25.

Botvinick, M., Jha, A. P., Bylsma, L. M., Fabian, S. A., Solomon, P. E., and Prkachin, K. M. (2005), "Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain." *Neuroimage*, 25, 312–19.

Clark, T. F., Winkielman, P., and McIntosh, D. N. (2008), "Autism and the extraction of emotion from briefly presented facial expressions: Stumbling at the first step of empathy." *Emotion*, 8(6), 803–9.

Cushman, F. (2008), "Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment." *Cognition*, 108(2), 353–80.

Cusi, A., Macqueen, G. M., and McKinnon, M. C. (2010), "Altered self-report of empathic responding in patients with bipolar disorder." *Psychiatry Research*, 178(2), 354–8.

Davis, M. H. (1983), "Measuring individual differences in empathy: Evidence for a multidimensional approach." *Journal of Personality and Social Psychology*, 44(1), 113–26.

De Greck, M., Wang, G., Yang, X., Wang, X., Northoff, G., and Han, S. (2012), "Neural substrates underlying intentional empathy." *Social Cognitive and Affective Neuroscience*, 7(2), 135–44.

Decety, J., Norman, G. J., Berntson, G. G., and Cacioppo, J. T. (2012), "A neurobehavioral evolutionary perspective on the mechanisms underlying empathy." *Progress in Neurobiology*, 98(1), 38–48.

Dennett, D. (1981a), "Intentional systems," in D. Dennett (ed.), *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge: MIT Press, pp. 3–22.

—(1981b), "Conditions of personhood," in D. Dennett (ed.), *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge: MIT Press, pp. 267–85.

—(1987), *The Intentional Stance*. Cambridge: MIT Press.

Dunfield, K. A. and Kuhlmeier, V. A. (2010), "Intention-mediated selective helping in infancy." *Psychological Science*, 21(4), 523–7.

Dziobek, I., Rogers, K., Fleck, S., Bahnemann, M., Heekeren, H. R., Wolf, O. T., and Convit, A. (2008), "Dissociation of cognitive and emotional empathy in adults with Asperger syndrome using the Multifaceted Empathy Test (MET)." *Journal of Autism and Developmental Disorders*, 38, 464–73.

Fiala, B., Arico, A., and Nichols, S. (2011), "On the psychological origins of dualism: Dual-process cognition and the explanatory gap," in E. Slingerland and M. Collard (eds), *Creating Consilience: Integrating the Sciences and the Humanities*. Oxford: Oxford University Press, pp. 88–110.

Frith, U. (2003), *Autism: Explaining the Enigma* (2nd edition). Oxford: Blackwell.

Gonzalez-Liencres, C., Shamay-Tsoory, S. G., and Brüne, M. (2013), "Towards a neuroscience of empathy: Ontogeny, phylogeny, brain mechanisms, context and psychopathology." *Neuroscience and Biobehavioural Reviews*, 37(8), 1537–48.

Gray, K. (2012), "The power of good intentions: Perceived benevolence soothes pain, increases pleasure, and improves taste." *Social Psychological and Personality Science*, 3(5), 639–45.

Gray, K. and Wegner, D. M. (2008), "The sting of intentional pain." *Psychological Science*, 19(12), 1260–62.

—(2009), "Moral typecasting: Divergent perceptions of moral agents and moral patients." *Journal of Personality and Social Psychology*, 96(3), 505–20.

—(2010), "Blaming god for our pain: Human suffering and the divine mind." *Personality and Social Psychology Review*, 14(1), 7–16.

—(2011a), "Morality takes two: Dyadic morality and mind perception," in M. Mikulincer and P. Shaver (eds), *The Social Psychology of Morality: Exploring the Causes of Good and Evil*. Washington, DC: American Psychological Association, pp. 109–27.

—(2011b), "To escape blame, don't be a hero—be a victim." *Journal of Experimental Social Psychology*, 47(2), 516–19.

Gray, H. M., Gray, K., and Wegner, D. M. (2007), "Dimensions of mind perception." *Science*, 315, 619.

Gray, K., Waytz, A., and Young, L. (2012a), "The moral dyad: A fundamental template unifying moral judgment." *Psychological Inquiry*, 23(2), 206–15.

Gray, K., Young, L., and Waytz, A. (2012b), "Mind perception is the essence of morality." *Psychological Inquiry*, 23(2), 101–24.

Gray, K., Jenkins, A. C., Heberlein, A. S., and Wegner, D. M. (2011), "Distortions of mind perception in psychopathology." *Proceedings of the National Academy of Sciences*, 108, 477–9.

Hamlin, J. K., Wynn, K., and Bloom, P. (2007), "Social evaluation by preverbal infants." *Nature*, 450, 557–9.

Harari, H., Shamay-Tsoory, S. G., Ravid, M., and Levkovitz, Y. (2010), "Double dissociation between cognitive and affective empathy in borderline personality disorder." *Psychiatry Research*, 175(3), 277–9.

Heberlein, A. S. and Adolphs, R. (2004), "Impaired spontaneous anthropomorphizing despite intact perception and social knowledge." *Proceedings of the National Academy of Sciences*, 101(19), 7487–91.

Heider, F. and Simmel M. (1944), "An experimental study of apparent behavior." *American Journal of Psychology*, 57, 243–59.

Huebner, B., Bruno, M., and Sarkissian, H. (2010), "What does the nation of China think about phenomenal states?" *Review of Philosophy and Psychology*, 1(2), 225–43.

Jack, A. I. and Robbins, P. (2012), "The phenomenal stance revisited." *Review of Philosophy and Psychology*, 3(3), 383–403.

Jackson, P. L., Rainville, P., and Decety, J. (2006), "To what extent do we share the pain of others? Insight from the neural bases of pain empathy." *Pain*, 125, 5–9.

Jones, A. P., Happé, F. G. E., Gilbert, F., Burnett, S., and Viding, E. (2010), "Feeling, caring, knowing: Different types of empathy deficit in boys with psychopathic tendencies and autism spectrum disorder." *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 51(11), 1188–97.

Kant, I. (1785/2005), *Groundwork for the Metaphysics of Morals*. Toronto: Broadview Press.

Kaplan, J. T. and Iacoboni, M. (2006), "Getting a grip on other minds: Mirror neurons, intention understanding, and cognitive empathy." *Social Neuroscience*, 1(3–4), 175–83.

Knobe, J. and Prinz, J. (2008), "Intuitions about consciousness: Experimental studies." *Phenomenology and the Cognitive Sciences*, 7, 67–83.

Lamm, C., Decety, J., and Singer, T. (2011), "Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain." *NeuroImage*, 54(3), 2492–502.

Lorenz, K. (1935), "Der Kumpan in der Umwelt des Vogels." *Journal für Ornithologie*, 83(2), 137–213.

Mier, D., Lis, S., Esslinger, C., Sauer, C., Hagenhoff, M., Ulferts, J., Gallhofer, B., and Kirsch, P. (2013), "Neuronal correlates of social cognition in borderline personality disorder." *Social Cognitive and Affective Neuroscience*, 8(5), 531–7.

Minzenberg, M. J., Poole, J. H., and Vinogradov, S. (2006), "Social-emotion recognition in borderline personality disorder." *Comprehensive Psychiatry*, 47(6), 468–74.

Mitchell, J. P., Macrae, C. N., and Banaji, M. R. (2006), "Dissociable medial prefrontal contributions to judgments of similar and dissimilar others." *Neuron*, 50(4), 655–63.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., and Gabrieli, J. D. (2011), "Impaired theory of mind for moral judgment in high-functioning autism." *Proceedings of the National Academy of Sciences*, 108(7), 2688–92.

Nummenmaa, L., Hirvonen, J., Parkkola, R., and Hietanen, J. K. (2008), "Is emotional contagion special? An fMRI study on neural systems for affective and cognitive empathy." *NeuroImage*, 43(3), 571–80.

Onishi, K. H. and Baillargeon, R. (2005), "Do 15-month-old infants understand false beliefs?" *Science*, 308, 255–8.

Paulhus, D. L., Hemphill, J., and Hare, R. (2009), *Manual of the Self-Report Psychopathy Scale* (SRP-III). Toronto: Multi-Health Systems.

Raine, A. and Benishay, D. (1995), "The SPQ-B: A brief screening instrument for schizotypal personality disorder." *Journal of Personality Disorders*, 9, 346–55.

Rankin, K. P., Kramer, J. H., and Miller, B. L. (2005), "Patterns of cognitive and emotional empathy in frontotemporal lobar degeneration." *Cognitive and Behavioral Neurology*, 18(1), 28–36.

Richell, R. A., Mitchell, D. G. V., Newman, C., Leonard, A., Baron-Cohen, S., and Blair, R. J. R. (2003), "Theory of mind and psychopathy: Can psychopathic individuals read the 'language of the eyes'?." *Neuropsychologia*, 41, 523–6.

Rizzolatti, G. (2005), "The mirror neuron system and its function in humans." *Anatomy and Embryology*, 210(5–6), 419–21.

Robbins, P. and Jack, A. I. (2006), "The phenomenal stance." *Philosophical Studies*, 127, 59–85.

Saxe, R. (2009), "The happiness of the fish: Evidence for a common theory of one's own and others' actions," in K. D. Markman, W. M. P. Klien, and J. A. Suhr (eds), *The Handbook of Imagination and Mental Simulation*. New York: Psychology Press, pp. 257–66.

Shamay-Tsoory, S. G., Aharon-Peretz, J., and Perry, D. (2009), "Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions." *Brain*, 132(3), 617–27.

Shamay-Tsoory, S., Harari, H., Szepsenwol, O., and Levkovitz, Y. (2009), "Evidence of impaired cognitive empathy in disorder." *Journal of Neuropsychiatry and Clinical Neuroscience*, 21(1), 59–67.

Silver, N. (2012), *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't*. New York, NY: The Penguin Press.

Singer, P. (1981), *The Expanding Circle: Ethics and Sociobiology*. New York, NY: Farrar, Straus & Giroux.

Singer, T. and Lamm, C. (2009), "The social neuroscience of empathy." *Annals of the New York Academy of Sciences*, 1156, 81–96.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., and Frith, C. D. (2004), "Empathy for pain involves the affective but not sensory components of pain." *Science*, 303, 1157–62.

Smith, A. (2006), "Cognitive empathy and emotional empathy in human behavior and evolution." *The Psychological Record*, 56, 3–21.

Sytsma, J. and Machery, E. (2009), "How to study folk intuitions about phenomenal consciousness." *Philosophical Psychology*, 22(1), 21–35.

—(2010), "Two conceptions of subjective experience." *Philosophical Studies*, 151(2), 299–327.

—(2012), "The two sources of moral standing." *Review of Philosophy and Psychology*, 3(3), 303–24.

Trivers, R. (1971), "The evolution of reciprocal altruism." *The Quarterly Review of Biology*, 46(1), 35–57.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J., and Cacioppo, J. T. (2010), "Making sense by making sentient: Effectance motivation increases anthropomorphism." *Journal of Personality and Social Psychology,* 99(8), 410–35.

Young, L., Koenigs, M., Kruepke, M., and Newman, J. P. (2012), "Psychopathy increases perceived moral permissibility of accidents." *Journal of Abnormal Psychology*, 121(3), 659–67.

6

More Than a Feeling: Counterintuitive Effects of Compassion on Moral Judgment

Anthony I. Jack, Philip Robbins,
Jared P. Friedman, and Chris D. Meyers

In a mechanistic civilization, there is grave danger of a crude utilitarianism, which sacrifices the whole aesthetic side of life to what is called 'efficiency.'

(Russell 1926, p. 31)

Emotions, even though their hallmark is the internal state of the individual—the viscera, the gut—are above all social phenomena. They are the basis of social interaction, they are the products of social interaction, their origins, and their currency.

(Zajonc 1998, pp. 619–20)

It is a dangerous error to confound truth with matter-of-fact. Our life is governed not only by facts, but by hopes; the kind of truthfulness which sees nothing but facts is a prison for the human spirit.

(Russell 1926, p. 129)[1]

1 Introduction

Morality lies at the heart of human social behavior, and emotions lie at the heart of social cognition. What then of the relationship between morality and emotion? An influential school of thought concerning this issue appears to have been guided by a simple model: there exists a fundamental tension between

reason and passion, and normatively correct moral judgment derives chiefly from the exercise of reason alone. We also think there is a fundamental tension between two kinds of psychological processes involved in moral cognition. However, we do not see the divide as being between reason and passion, but rather between two forms of reason: one which is emotionally detached, analytic, and logical in nature, and a second which does deal with emotions as one of its primary currencies, yet which is more akin to a refined aesthetic. In other words, the key divide is not so much between reason and passion as it is between sense and sensibility, where sensibility is a partly affective and partly cognitive affair—that is, more an active mode of understanding than a passive mode of feeling. On this view, it is not our capacity for detached analytic thought which lies at the heart of moral understanding, but rather our social and emotional sensibility. This is not to deny that our capacities for logical and scientific thought serve as necessary adjuncts to moral reasoning; at the very least they are essential for us to understand the landscape in which our moral judgments must take shape. But these capacities also pose a risk to moral reason. When we get caught up in an empirical worldview, it is all too easy for us to lose sight of what really matters. Analytic thinking, we suggest, can cause us to lose sight of our humanity.

The key to determining which of these two views is more plausible—reason versus passion, or sense versus sensibility—lies in understanding the role that emotions play in moral judgment. This is the subject of this chapter. Experimental psychologists and neuroscientists have focused extensively on a class of moral dilemmas that have been explicitly designed to pit reason and emotion against each other. While it is clear from prior work that both emotion and reasoning are elicited by these scenarios, it is less clear which emotions are implicated, as well as how and why those emotions influence moral judgment. Accounts of moral cognition based on classic dual-process theory hold that these emotional processes are primitive, automatic, and cognitively shallow in nature. We begin here by motivating an alternative view, which derives from combining our own philosophically inspired theory of cognitive structure with recent evidence from cognitive neuroscience. We then present evidence from five experiments which support our model—evidence that the role of emotion in moral judgment is more nuanced and cognitively complex than can be readily accommodated by accounts based on classic dual-process theory.

1.1 A moral tension

Contemporary ethical theory is dominated by two perspectives on the nature of right action. According to utilitarianism, right actions are those actions that maximize aggregate happiness (Mill 1861/1998). According to deontological ethics, the moral status of an action depends on whether or not the action conforms with an abstract rule, such as the injunction against treating others merely as means to an end, as opposed to ends in themselves (Kant 1785/2005). The contrast between these views is best exemplified by cases of moral decision-making that promote conflicting verdicts about the right thing to do. In paradigm cases of this type, picking the utility-maximizing option means neglecting the duty to respect basic individual rights, such as the right to life. A familiar example of this is the "footbridge" variant of the trolley problem, in which saving the lives of a group of strangers requires pushing an innocent bystander to his death (Thomson 1985). This is just one example of a large class of hypothetical scenarios, all of which share a common causal-intentional structure, and all of which present stark, life-and-death choices: you can either promote the general good by sacrificing an individual, or you can forgo that sacrifice at the expense of the general good. You can't have it both ways.

Informed by this philosophical background, neuroscientific and psychological research on morality has also focused on the contrast between utilitarian and deontological perspectives. For example, Greene and colleagues have assembled a neuroscientific account of moral judgment which borrows heavily from classic dual-process theory (Kahneman 2003). Greene has shown that different brain regions are recruited for different types of moral judgment. In particular, lateral regions associated with problem solving and executive functions tend to be recruited more for reasoning about scenarios that do not involve a conflict between utilitarian and deontological principles, whereas medial regions associated with emotional processing tend to be recruited more for reasoning about footbridge-type dilemmas, in which utilitarian and deontological perspectives clash (Greene et al. 2001; Greene et al. 2004). According to Greene's view, utilitarian choices reflect "cool," controlled, and analytic processing, whereas deontological choices are emotion driven, primitive, automatic, and intuitive (Greene 2007, 2012; Greene et al. 2001;

Greene et al. 2008; Greene et al. 2009; Cushman and Greene 2011). Greene sums up the view as follows:

> Deontological judgments tend to be driven by emotional responses and . . . deontological philosophy, rather than being grounded in moral *reasoning*, is to a large extent an exercise in moral *rationalization*. This is in contrast to consequentialism, which arises from rather different psychological processes, ones that are more 'cognitive,' and more likely to involve genuine moral reasoning. (Greene 2007, p. 36; emphasis in original)

1.2 Motivating an alternative perspective

Before we get to the differences, it will be useful to clarify the points of similarity between our account and Greene's. First, we have also arrived at a dual-process theory of moral judgment. We are in agreement with Greene that there exists a tension between two basic types of cognitive process which often guide moral judgments. Second, we agree with Greene about the neuroanatomical basis of these dueling influences on moral judgment. We have conducted our own neuroimaging research which uses different tasks to identify brain networks similar to and overlapping those that Greene has identified (Jack et al. 2013a; Jack et al. 2013b). Our studies use additional methods to demonstrate that the tension between these networks is a fundamental feature of our neural structure, and is not specific to the domain of moral judgments. Third, we agree with Greene that one of these networks is involved in various forms of nonsocial reasoning, and that emotions play a more significant role in the other network.

While these points of agreement are notable, our account also differs from Greene's in some important respects. These differences will be best understood by appreciating the different origins of the two accounts. Greene's account primarily borrows from classic dual-process theory and applies that model to moral judgment. Classic dual-process theory not only guides his model of moral cognition, but also his functional account of the regions identified by the neuroimaging of morality. By contrast, our account has been developed independently of classic dual-process theory and research on moral neuroscience—we are exploring those links for the first time here. Instead, our account derives from our own theory of cognitive structure (Robbins and Jack 2006; Jack and Robbins 2012; Jack forthcoming). It is a modification

of a theory implicit in Dennett's work and was originally conceived to provide a cognitive account of the origins of the philosophical problem of consciousness. According to our original theoretical conception (Robbins and Jack 2006), there are dissociable brain networks specialized for thinking about an entity's physical (e.g., causal-mechanical) attributes and for thinking about an entity's subjective mental life. The first network is activated when we take up the "physical stance" toward the entity, the second when we take up the "phenomenal stance." The network architecture is configured in such a way that the engagement of one stance inhibits the engagement of the other. Alongside these mutually antagonistic stances is a third stance that is compatible with both, namely, the "intentional stance," which is engaged when reasoning about an entity's goal-directed behaviors. Of these three stances, we submit that one of them in particular—the phenomenal stance—plays a distinctive role in moral reasoning, as this is the stance we step into when thinking about an entity as a target of moral concern. Each of these stances is associated with a distinct neuropsychological profile: Williams syndrome represents a deficit which predominantly affects the physical stance; autism spectrum disorders predominantly involve dysfunction of the intentional stance; and psychopathy, particularly the primary psychopathic trait of callous affect (i.e., lack of empathic concern), involves dysfunction of the phenomenal stance (see Figure 6.1).

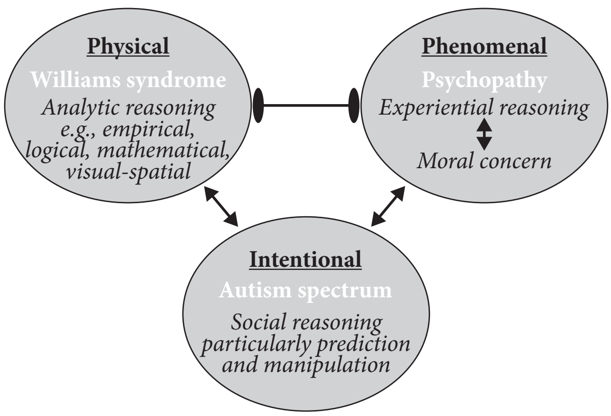


Figure 6.1 Our theoretical model, outlining three stances (or cognitive modes), their relationships, and disorders that affect them. (*Arrows* indicate mutual compatibility; *barbell* indicates mutual antagonism.)

Since its initial formulation, our theory has matured in concert with data from cognitive neuroscience. We have mapped the physical and the phenomenal stances onto distinct brain networks, which correspond to the networks commonly known as the "task-positive" and "default mode" networks, respectively (Jack et al. 2013a; Jack et al. 2013b). The task-positive and default mode networks have previously been shown to be in a natural state of tension, that is, their activity is negatively correlated or "anti-correlated" even when participants are not engaged in any explicit cognitive task (Fox et al. 2005). In other words, even when participants are at rest, activity in these networks tends to alternate much like a seesaw. We have shown that tasks involving scientific reasoning, and thinking about the internal mental states of others, push this seesaw to extremes. When we think about experiential mental states, we activate the default mode network and suppress the task-positive network; when we engage in scientific reasoning, we activate the task-positive network and suppress the default mode network (Jack et al. 2013a). The intentional stance, on the other hand, appears to recruit regions from both networks, breaking with the predominant tendency for one network to be suppressed when the other is activated.

Our original theory has been updated in line with emerging evidence from cognitive neuroscience, which suggests an evolutionary basis and a broader functional characterization of the two networks (Jack forthcoming). We hypothesize that we evolved distinct networks to do the cognitive processing required to guide distinct types of interaction: manipulating inanimate objects and engaging with conscious agents. Two broad cognitive modes correspond to these two types of interaction: analytic thinking and empathetic engagement. The first cognitive mode, which includes logical, mathematical, and causal-mechanical reasoning, was built upon our more primitive capacities for sensory processing and the control of action. Because it is built upon evidence from the senses, it can be thought of as an empirical mode of thinking. The second cognitive mode, which plays a key role in social bonding, moral cognition, introspection, and emotional insight, was built upon our more primitive capacities for visceral awareness and emotional self-regulation. This second cognitive mode is the default mode for unguarded social interactions, in particular between a parent and child but also more generally for in-group members—in other words, for anyone whom we humanize (Jack et al. 2013b).

However, it is not the only mode for social interaction. A more emotionally distanced mode of social interaction, corresponding to the intentional stance, involves a blend of these two cognitive modes which are naturally opposed to one another. This blending of the two cognitive modes is reflected in ordinary language. When we refer to someone as "calculating" or "manipulative," we do not literally mean that they are doing sums or using their fine motor skills; rather, we are referring to an emotionally distanced mode of social cognition. We likely use these terms because this mode of social cognition involves the same brain areas associated with mathematical calculation and fine motor control. Hence, for instance, when conditions involving deception are compared with nondeceptive conditions, differences in brain activity are seen in areas associated with analytic thinking (Christ et al. 2009). More Machiavellian individuals also activate this network more when engaged in social cognition (Bagozzi et al. 2013). Finally, this is the pattern that is evident when participants view dehumanizing narratives that depict others as subhuman animals (Jack et al. 2013b).

1.3 Morality and emotion

How does our model relate to moral cognition? We adopt a dynamic view of moral decision-making. In our view, good judgments involve considering ethical issues from multiple stances, then balancing these alternate perspectives against each other. Utilitarian thinking clearly has allegiance to the analytic mode of cognition. After all, utilitarian thinking involves above all else a calculus. It is true that it also involves consideration of psychological states (i.e., happiness vs. suffering), however, the calculation is based on an emotionally distanced form of reasoning about these states (which must somehow be quantified). Hence utilitarian reasoning would appear to rely predominantly on the physical and intentional stances.

We believe that utilitarian reasoning can play a useful role in moral decision-making. However, our view is that moral reasoning essentially involves the phenomenal stance. The personal engagement engendered by stepping into the phenomenal stance is associated with specific social emotions. Individuals that we include in our moral circle tend to provoke emotions such as compassion and resentment (Strawson 1962).

This, we contend, is a crucial aspect of moral reasoning. That is, our view is that we can only be said to be fully morally engaged with others if we are willing to delve into and attempt to understand their experiential perspective[2]; and our model holds that understanding the experiences of others is inextricably linked to moral concern (Figure 6.1). Regardless of whether we approve or disapprove of an individual's actions, this interpersonal engagement (sometimes called "intersubjectivity") appears to be essential for genuine moral understanding. Our view is that Kant's work on deontological reasoning at least partially captures the kinds of considerations and principles that emerge as central when we engage in this form of intersubjective reasoning.

Kant made a thorough-going distinction between moral reasoning and emotion. However, putting this difference aside, we hold quite consistent views concerning the appropriate role of emotion in a morally virtuous life. In particular, we agree with Kant that good moral decisions are not "emotion driven." Kant identified two types of feeling, referred to as "Affecten" (translated as "affects") and "Leidenschaften" (translated as "passions"), which are hindrances to good moral reasoning and which we have a duty to try to control (Formosa 2011). On the other hand, Kant thought we should expose ourselves to the sick and needy in order to "cultivate the compassionate natural (aesthetic) feeling in us" (Kant 1797/1996, 6:457). We agree with Kant that good moral reasoning requires us to "get hold of" and cultivate our emotional responses. For Kant, the purpose of this is to align our emotions with moral reason. Our view is different—we think that moral reasoning is inextricably linked to specific emotions. Hence, we hold that activation of brain regions involved in moral reasoning should be positively associated with the other-directed emotion of compassion, but that these regions are involved in the regulation of self-directed emotions such as personal distress, rather than being "emotion driven."

1.4 Key differences between the models and neuroimaging evidence

According to our model, the phenomenal stance is the cognitive mode that allows us to understand and make sense of both our own and other people's

experience. We do not regard this as a primitive, effortless, or cognitively trivial process. We believe it is a highly evolved capacity: although it is probably not uniquely human (De Waal 2009), it appears to be much more developed in the human than other animals. We also believe this capacity can be trained, much as our analytic[3] reasoning skills can be trained. Finally, we believe that a deep appreciation of experience requires sustained attention and reflection. All of these facets are reflected by the uniquely human tendency to create and consume complex works of art, music, and literature, activities which appear to centrally involve the act of reflecting upon the nature and meaning of human experience. Because we view this type of thinking as sophisticated, educable, and reflective, we believe it is appropriately characterized as a type of reason. This gives rise to a distinctive form of dual-process theory, which we will refer to as opposing domains theory.

Opposing domains theory holds there is a fundamental tension between two types of conscious, evolutionarily advanced, nonautomatic (controlled or reflective) reason: analytic versus empathetic reasoning. Opposing domains theory may be consistent with classic dual-process theory—we are open to the view that there is an orthogonal division between automatic and reflective reasoning processes, in addition to our claim that there is a division between two types of reflective reasoning. The accounts need only find themselves in direct competition when they both seek to explain a particular phenomenon. Opposing domains theory is in direct competition with Greene's dual-process account of the processes that push for and against utilitarian moral judgments.

Greene characterizes the processes that push us away from utilitarian moral judgments, and correspondingly the functions of the brain areas which we identify with the phenomenal stance, as emotion driven, automatic, and primitive. These are characteristics of Type 1 processes evoked in a number of classic dual-process accounts, although it is also important to note there is considerable variation and inconsistency between different dual-process accounts (Evans and Frankish 2009). In informal descriptions of his model, Greene consistently emphasizes the emotional nature of these processes and also describes them, metaphorically, as being similar to the automatic setting on a camera (Greene 2010), or as reflecting the workings of our "inner chimp" mind (Radiolab 2007).

To summarize, both models agree that there is one set of "cognitive" or "analytic" processes which are involved in a variety of nonsocial reasoning tasks. The key differences concern the characterization of the second set of processes. For Greene, these processes are (1) essentially emotional in nature. For us, emotion processing represents one important facet of these processes, but it does not define them. For Greene, these processes are (2) automatic and hence "emotion driven." For us they are nonautomatic and deliberative, and better characterized as being involved in emotion regulation. For Greene, these processes are (3) primitive. For us, they are evolutionarily advanced. We also differ in the link we suppose to exist between these two sets of processes and moral reasoning. We both agree that moral judgments are frequently influenced by both sets of processes. For Greene, (4) moral *reasoning* (contrast with moral intuition) involves the first ("cognitive" or "analytic") set of processes. For us, *moral* reasoning (contrast with analytic reasoning) involves the second ("empathetic") set of processes.

Both models borrow heavily from neuroimaging research and make claims about the same brain regions. Greene's early and groundbreaking work on the neural basis of moral cognition first appeared before the fields of social and affective neuroscience took off. When Greene got started (Greene et al. 2001) it was plausible enough to characterize the two brain networks as being involved in reason on the one hand and emotion processing on the other. More recent evidence, however, suggests that this model isn't entirely accurate. Figure 6.2 summarizes some key points.

All the panels in Figure 6.2 show the borders of regions defined by a formal metaanalysis of studies of moral cognition (Figure 6.2A). This meta-analysis and others shown in the figure derive from an extensive study by our laboratory which is currently in preparation (Tillem and Jack 2013). Notably, our morality metaanalysis, which includes a number of studies of Greene's as well studies from other groups, identifies a relatively delimited set of regions as being consistently and robustly associated with moral cognition (as compared to control conditions). There is little consistent activation in regions classically associated with nonsocial reasoning; however, there is consistent activation in three regions which have been repeatedly identified with social cognition (Amodio and Frith 2006; Mitchell 2009; Van Overwalle 2009). These are labeled in the first panel, and consist of medial prefrontal cortex (MPFC), medial

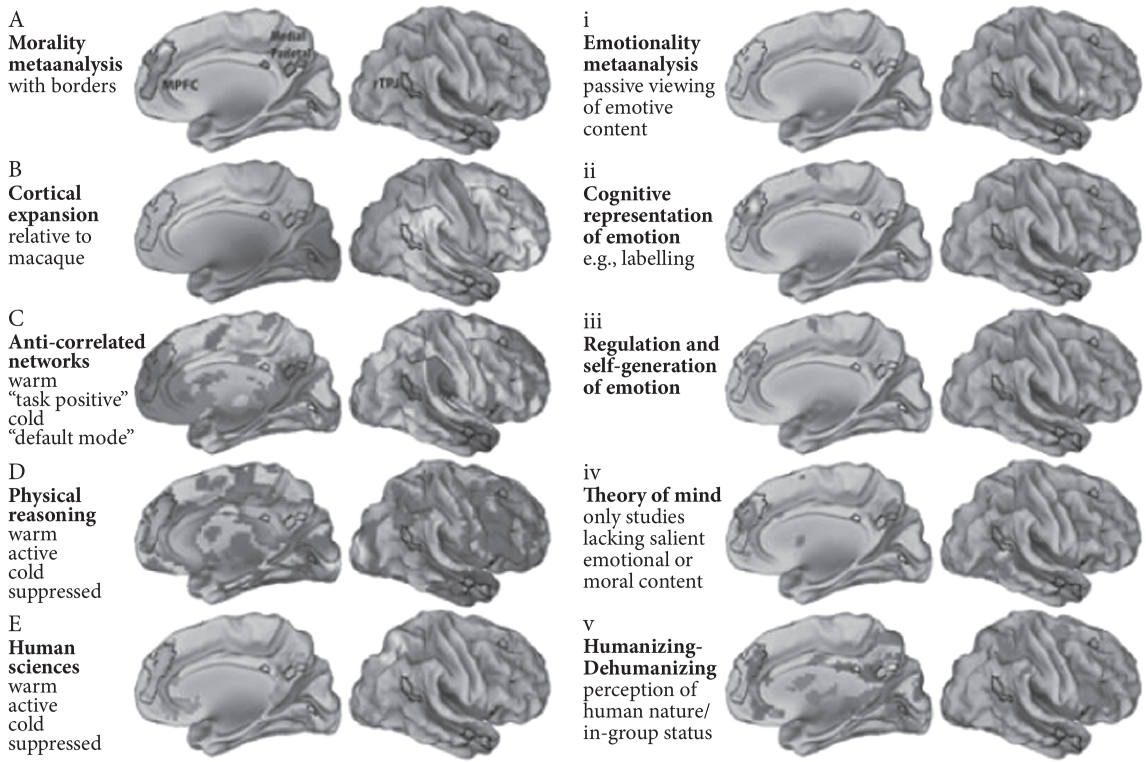


Figure 6.2 Metaanalysis of moral reasoning areas and comparison with other findings. Panels on the left (A–E) address issues about the general properties of regions involved in moral cognition. Panels on the right (i–v) address issues relating to their active functional role. A color version of this figure is available online at: http://www.bloomsbury.com/advances-in-experimental-philosophy-of-mind-9781472514806/

parietal cortex, and right temporo-parietal junction (rTPJ). The latter region appears to be most strongly involved in the representation of the beliefs and intentions of others (Saxe and Kanwisher 2003), including in moral reasoning (Young et al. 2010). Of greatest relevance are the two midline regions, MPFC and medial parietal, which are consistently highlighted in reviews of moral reasoning (Moll and de Oliveira-Souza 2007; Greene 2009), and which Greene labels as emotion areas.

The first observation we make is that moral judgment tasks consistently and robustly activate empathetic brain regions, but not analytic brain regions (Figure 6.2A). This fits well with our view that moral reasoning essentially involves these regions. At first sight, this observation might raise some concern about Greene's view that moral reasoning involves analytic brain areas. However, it does not contradict his view. Much research in the classic dual-process tradition suggests that people often forgo analytic reasoning even when it is clearly more appropriate for the task at hand. Similarly, Greene holds that people typically forgo moral reasoning, and instead rely on moral intuitions (Greene 2010). This is a pessimistic view, and not the view we hold, but it is consistent with the evidence.[4]

The second observation relates to the claim that these regions are primitive. Figure 6.2B shows a cortical expansion map derived by comparing landmarks on the human cortical surface to homologous regions in the macaque monkey (Van Essen and Dierker 2007). It is clear from this map that moral reasoning does not recruit the most expanded areas. The brightest regions correspond to areas some 32 times larger in the human than the macaque. In fact, the areas that demonstrate the greatest expansion are involved in analytic reasoning and correspond quite closely to the task-positive network (Figure 6.2C). However, the regions involved in moral reasoning are clearly highly expanded, being approximately 20 times greater in the human than the macaque. Other work similarly supports the view that these regions have undergone major expansion (Hill et al. 2010). Hence the label "primitive" is not accurate. Areas which can be more appropriately labeled primitive appear in darker red on the map. They lie in early visual cortex, primary motor cortex, and subcortex.

Third, Figure 6.2C illustrates an important advance on Greene's early and insightful interpretation of these medial moral regions as being in tension with

analytic reasoning areas during moral reasoning. We now know that these regions are constantly in tension with analytic reasoning areas, not just for moral reasoning tasks, but in a variety of other experimental tasks and also in the absence of any task. This panel illustrates two networks whose activity tends to be inversely correlated. The darker colors highlight the regions that show the greatest tension. Warmer colors correspond to the task-positive network, a network which is recruited during a wide variety of nonsocial tasks (Raichle et al. 2001), but which is suppressed during empathetic social cognition (Jack et al. 2013a). The colder colors correspond to the default mode network, which is suppressed during a wide variety of nonsocial tasks, but activated above baseline during empathetic social cognition. The tension between these networks, which is a highly pronounced feature of our neurophysiology, provides the neuroscientific basis for opposing domains theory. Panels D and E illustrate two specific types of task which tend to suppress the default mode network, including the medial moral regions. Panel D illustrates that these regions are suppressed when participants solve physics problems (Jack et al. 2013a). Panel E illustrates that these regions are suppressed when participants read descriptions taken from the human sciences (biology, neuroscience, and psychology; see Jack et al. 2013b). These illustrations are significant when one considers that utilitarianism represents a scientific approach to morality, in that it represents an attempt to systematize moral judgments by adopting a quantitative approach.

An important implication of the findings summarized in panels C–E is that these medial moral regions are not engaged in automatic processing. Activity in these regions is suppressed in proportion to cognitive load (Raichle et al. 2001), a hallmark of controlled processing. The same effect of load has been observed in a dual-task situation where participants were engaged in mentalizing (Spunt and Lieberman 2013).

Next, let us turn to the right-hand panels which shed light on the positive functions of these regions. Greene has characterized the midline moral regions (MPFC and medial parietal) as being involved in automatic, emotion-driven processing. The first panel shows brain regions consistently recruited by tasks that involve the passive viewing of emotionally arousing content. These tasks recruit subcortical structures, such as the amygdala and insula, activation of which is associated with being in the grips of an emotional state.

However, they do not recruit the MPFC or medial parietal regions. This finding is inconsistent with a view that characterizes these regions as automatic or emotion driven. A more widely held view among affective neuroscientists is that the MPFC is involved in the cognitive representation of emotions in both self and others, as well as emotion regulation (Ochsner et al. 2004). These points are illustrated by our metaanalytic findings shown in panels (ii) and (iii). Panel (iii) illustrates that the MPFC is not so much emotion driven as it is involved in the top-down regulation of emotional states (e.g., where participants are explicitly instructed to down-regulate, up-regulate, or generate emotions). This fits our model, and is consistent with Kant's view that moral reasoning requires us to "get hold of" certain emotion-driven responses.

Panel (iv) shows that both MPFC and medial parietal are reliably recruited by theory of mind tasks which lack any obvious emotional or moral content. Theory of mind tasks are not well characterized as emotion driven, but are commonly held to involve a type of social reasoning. In sum, only one of the two "emotion" regions identified by Greene has any clear involvement in emotion processing; the role it plays in emotion processing is quite different from how Greene has characterized it, and both regions are involved in a type of reasoning.

This review indicates that, with regard to the midline brain areas identified as being involved in emotion by Greene et al. (2001), the evidence favors our account over Greene's for each of the four key differences between the accounts identified above.

Can anything more be said about the role of these regions in moral cognition, and the tension that is seen between them and analytic reasoning areas? Panel (v) shows the results of a conjunction analysis from our recent study on the perception of humanness (Jack et al. 2013b). This most clearly implicates the medial parietal cortex. We found this area was more active when participants read humanizing as opposed to two different types of dehumanizing narrative: mechanistic dehumanizing (or objectifying) and animalistic dehumanizing. We also found it was more active when people viewed pictures of unfamiliar human faces, as opposed to pictures of either animals or machines. This area is even more active when participants view pictures of known individuals, as contrasted to unfamiliar individuals (Gobbini and Haxby 2007). Hence, this

region appears to be highly sensitive to the degree to which the object of our attention lies within our moral circle. This region is also the central node of the default mode network, and shows the strongest anti-correlations with analytic brain areas (Fransson and Marrelec 2008). The blue areas in panel (v) show that activity in analytic brain areas is inversely associated with the perception of humanness. Hence, it appears that analytic reasoning is associated with a tendency to dehumanize others—to regard them instrumentally, as mere objects, rather than as ends-in-themselves.

1.5 Neuroimaging of individual differences

So far we have focused on brain areas that are found to be more active for personal ("high conflict") than impersonal moral dilemmas. According to Greene, these areas are generally involved in emotion processing, whereas according to our model they are more involved in social reasoning, in particular understanding the experiences of others. We identify compassion, or empathic concern, as one emotion which is central to the function of these brain areas, because compassion centrally involves appreciation of, and concern about, the negative experiences of others. It is well established that participants tend to make more utilitarian judgments when faced with impersonal moral dilemmas (i.e., diverting the trolley so that one person dies rather than five) rather than personal moral dilemmas (i.e., directly harming one person to save five others). In addition, work on objectification and dehumanizing (Cikara et al. 2011; Harris and Fiske 2006; Jack et al. 2013b) establishes a link between compassion and activity in these brain areas. However, there is also another way to examine the link between utilitarianism and compassion. There is individual variation in the degree to which people choose the utilitarian option for personal moral dilemmas. In this section, we extend our review from the previous section to look at work that sheds light on the brain areas associated with these individual differences.

Imaging work that looks specifically at the personal moral dilemmas indicates that activity in one area in particular, comprising the ventromedial prefrontal cortex (vMPFC) and the neighboring subgenual cingulate, appears to be critical to whether participants pick or resist the utilitarian choice (Wiech et al. 2013). This area lies ventral to (i.e., below) the MPFC

region identified in Figure 6.2. It is implicated in our study of the perception of humanness (Figure 6.2, panel v), and it has been directly implicated in empathic concern for others (Zahn et al. 2009). However, this region is not clearly involved in social reasoning of the type required by theory of mind tasks (Figure 6.2, panel iv). Instead, it is seen in mental time travel where one projects oneself into the past or future (Buckner and Carol 2007). Although there is clearly a different story to be told about the function of this area, it has many of the same properties which critically differentiate between our account and Greene's for the MPFC and medial parietal areas. The vMPFC is part of the default network and is typically suppressed when participants engage in demanding analytic/nonsocial tasks (Raichle et al. 2001), notably including descriptions from the human sciences (Figure 6.2, panel E). The vMPFC is not automatically activated by emotional stimuli. However, it is linked to a distinct form of emotion regulation associated with a sense of social connectedness and the parasympathetic nervous system. Activity in this region mediates the mitigation of feelings of distress that occur when participants think about a loved one (Eisenberger and Cole 2012). Hence, this region is not involved in automatic, emotion driven responses. Nor is this region "primitive"—it has undergone considerable evolutionary expansion. While this brain area is not involved in either social reasoning or affect per se, it does appear to be involved in a type of affective representation—namely, determining the subjective value of objects and outcomes. Hence, it has been suggested that it plays a key role in integrating affective and conceptual information to generate a sense of affective meaning (Roy et al. 2012).

Again, this evidence suggests that Greene's model does not do justice to the complex and multifaceted role of the emotions in moral reasoning. In particular, it glosses over the distinctive contribution of prosocial emotions such as compassion or empathic concern, understood as an other-oriented affective response to the plight of another person (Batson 2009). It does not fit well with what is known about these prosocial emotions to characterize the judgments and choices influenced by them as emotion driven. While fear and anger are associated with the sympathetic nervous system and its "fight or flight" response, prosocial emotions are associated with the parasympathetic nervous system, which modulates sympathetic nervous

system response. This suggests that reasoning based on prosocial emotions is not so much emotion driven as it is associated with a specific form of emotion regulation. Further, it is widely recognized that both prosocial emotions and certain aspects of moral judgment depend upon complex cognitive representations, in particular representations of the mental states of conspecifics (Haslam 2006; Smith 2011; Gray et al. 2012; Cushman and Young 2011). This view fits well with the evidence that has emerged from social neuroscience about the broader functional role of the "emotional processing" areas identified in early work by Greene. It fits poorly with Greene's view of the effects of emotion on moral judgment as primitive and divorced from reason.

1.6 From neuroimaging to behavior

The neuroimaging evidence we have reviewed is more consistent with our account than Greene's. In this section we turn to behavioral evidence.

There is now a substantial body of evidence that directly links individual differences in prosocial emotion to resisting the utilitarian option for personal moral dilemmas. For example, the tendency to favor the utilitarian choice in footbridge-type scenarios has been shown to correlate positively with antisocial personality traits in the normal population (Bartels and Pizarro 2011), as well as with psychopathy (Koenigs et al. 2012), acquired sociopathy (Koenigs et al. 2007), and a genetic variation that predisposes individuals to psychopathy (Marsh et al. 2011; Glenn 2011). These findings suggest that prosocial personality traits are likely to correlate with resisting the utilitarian option. Consistent with this hypothesis, Côté et al. (2013) show that the positive correlation between social class and utilitarian judgment in a resource allocation task (i.e., higher-class individuals lean more utilitarian) is mediated by empathic concern (i.e., it occurs because higher-class individuals evidence less empathic concern). More directly, Choe and Min (2011) show that utilitarian judgment in footbridge-type dilemmas is negatively correlated with empathy, as measured by the Emotional Empathic Tendency Scale (Meharabian et al. 1988), of which empathic concern is a component (for similar findings, see Conway and Gawronski 2013). More recently, Gleichgerrcht and Young (2013) showed that diminished levels of empathic

concern predicted utilitarian judgments to personal-harm dilemmas, even when controlling for demographics, religiosity, moral knowledge, and independent psychological constructs of empathy (i.e., perspective taking, personal distress, and fantasy).

This work establishes that a highly robust difference between utilitarian and nonutilitarian responders on personal moral dilemmas is that the former are less able to feel concern for the suffering of others. This difference can be interpreted in two ways. First, it might be read as broadly supporting Greene's model that resistance to utilitarian responding is emotion driven. Indeed, Greene (2009) has cited some of this work in this context. On the other hand, the link also provides support for the more specific prediction made by our model, which identifies moral concern in particular as central to the phenomenal stance and resistance to utilitarian thinking. Given the equivocal nature of this evidence in relation to the two models, further evidence is needed to determine which model is better supported.

In the experiments we report here, we present further evidence for our model. First, we establish that the link between emotionality and resistance to utilitarianism is specific to prosocial emotion; it does not hold for other measures of emotionality. Second, we show the effects of prosocial emotion are mediated by our understanding of the minds of the targets involved, in that the effect of prosocial emotion is present for humans but absent for animals. This provides further evidence that this effect is not simply "emotion driven" but instead represents an integration of affective and conceptual information. Third, we examine a novel, and more "real life," type of case which pits utilitarian reasoning against other concerns—euthanasia. This case reveals a highly counterintuitive finding: although measures of compassion directly assess sensitivity to the suffering of others, individuals with greater compassion are in fact more willing to allow others to suffer than to allow them to be directly harmed. Again this effect is present for humans but actually reverses for subhuman animals. This surprising finding suggests something more than the mere mediation of prosocial emotion by mind representation. It suggests that prosocial emotion influences our conception of others, causing us to attribute to them, and value, mental properties which go beyond the basic affective experiences that are the focus of a utilitarian calculus.

1.7 Overview of behavioral experiments

Study 1 demonstrates that the link between individual differences in emotionality and utilitarian reasoning is specific to prosocial emotion; other measures of emotionality do not predict different responses to the standard dilemmas. Study 2 demonstrates that the link between prosocial emotion and moral reasoning is specific to a particular sort of deontological principle, namely, the prohibition against directly harming a person. There is no link with prosocial emotionality for scenarios in which the utilitarian choice requires lying, cheating, breaking promises, or otherwise violating norms of justice and fairness. These first two studies replicate and extend the findings reported by Gleichgerrcht and Young (2013).

Study 3 establishes that the link between prosocial emotionality and prohibitions against harming are even more specific: the link only holds for cases involving humans, and not for matched cases involving animals. This suggests that the effects of prosocial emotion on moral reasoning are contingent on how we represent the minds of those involved. Studies 4 and 5 shed further light on this by examining a different type of case involving different sorts of harm to a single individual. Euthanasia judgments pit the utilitarian concern to minimize suffering against the deontological concern to preserve and protect life, even a life of suffering. Again we find that utilitarian judgments are negatively correlated with compassion only for targets that are humans, not for animals. While prosocial emotion has typically been primarily associated with concern for the suffering of others, and is measured by questions designed to assess that sentiment, our findings suggest that an association between prosocial emotion and respect for human life is the more powerful factor driving moral judgment. Hence, it appears that prosocial emotions can, surprisingly, motivate behavior that has the (undesired) consequence of prolonging others' suffering.[5]

In summary, these findings suggest that prosocial emotions influence moral reasoning first and foremost by increasing the individual's tendency to reason in a manner that accords with an abstract moral principle. This emergent moral principle does not apply to minds in general, but is specific to human(-like) minds. Hence, it appears that prosocial emotions have a highly nuanced effect on moral judgments which is tightly linked to how we represent the minds of those involved. In short, while it may be partially correct to say

that emotion drives deontological moral judgment—and we acknowledge this core insight in Greene's work—it is apparent on closer examination that the nature of that influence is specific to prosocial emotion and indistinguishable from a form of reason.

2 Study 1

Though the idea that emotions play a role in moral decision-making is relatively uncontroversial, the nature of that role is not well understood. According to Greene's model, resistance to the utilitarian option in footbridge-type scenarios is driven by rapid-fire, "alarm-like" emotional responses to the thought of directly harming an individual; however, the model does not specify the character of the emotional component in any further detail. Our first experiment represents an initial attempt to remedy this lack of specificity in the model. To that end, we combine a standard measure of utilitarian judgment with a battery of measures of social-emotional functioning, including empathic concern, emotionality, emotion regulation, emotional self-awareness, antisociality, social intelligence, and the tendency to override intuitive responses. We included this suite of measurements to test the hypothesis that an enhanced capacity for empathetic engagement in particular, rather than a high level of emotionality generally, would exhibit a selectively negative relationship with utilitarian reasoning. In other words, we predicted that measurements of prosocial and antisocial tendencies would exhibit a unique relationship to higher and lower levels of deontological reasoning, respectively, relative to other measures.

2.1 Participants and procedure

Participants were 53 undergraduate students (mean age 20.2 years) who took part in lab-administered tests at Case Western Reserve University. This study (like the other studies reported in this chapter) was approved by the Institutional Review Board on campus prior to data collection.

Participants first completed individual difference measures from several categories, summarized in Table 6.1:

Table 6.1 Individual difference measures, by category. (Abbreviations: IRI = Interpersonal Reactivity Index; SRP-III = Self-Report Psychopathy Scale, version 3.)

Category	Scale
Prosociality	IRI Empathic Concern (Davis 1980)
Emotionality	Berkeley Expressivity Questionnaire (Gross and John 1997); IRI Personal Distress
Emotion regulation	Emotion Regulation Questionnaire (Gross and John 2003)
Self-awareness	Five Facet Mindfulness Questionnaire (Baer et al. 2006)
Intuition suppression	Cognitive Reflection Test (Frederick 2005)
Antisociality	Aggression Questionnaire (Buss and Perry 1992); SRP-III Callous Affect, SRP-III Interpersonal Manipulation (Paulhus et al., forthcoming)
Social intelligence	IRI Perspective Taking; Social Stories Questionnaire (Lawson et al. 2004); Imposing Memory Task, theory of mind subscore (Kinderman et al. 1998); Reading the Mind in the Eyes Task, Revised Edition (Baron-Cohen et al. 2001); Diagnostic Analysis of Nonverbal Accuracy, facial emotion recognition task (Nowicki and Duke 2001)

Self-report measures of social-emotional functioning (Interpersonal Reactivity Index, Self-Report Psychopathy Scale, Emotion Regulation Questionnaire, Five Facet Mindfulness Questionnaire, Aggression Questionnaire) were completed using online surveys, which participants filled out in their own time between laboratory testing sessions. In addition to self-report measures, peer measures of empathy, emotional self-control, and emotional self-awareness were taken by asking participants to supply the email addresses of friends, who were then contacted by email and asked to fill out a confidential survey about the participant. The survey in question included subscales of the Emotional and Social Competency Inventory, University Edition (Boyatzis and Goleman 2007), including the six-item Empathy subscale (sample items of which include "Understands others from different backgrounds" and "Understands others by putting self into other's shoes"). Individual items were scored on a 7-point scale (1 = Never, 7 = Always). A peer rating was generated for each participant by averaging across peers. Non-self-report measures of social intelligence (Social Stories Questionnaire, Imposing Memory Task, Reading the Mind in the Eyes Task, DANVA facial emotion recognition task) were completed in the

laboratory on a computer, using E-Prime 2.0 software. Participants also took a computerized version of the Cognitive Reflection Test, which assesses the ability to override intuitive (but incorrect) responses to mathematical "brain teasers" in favor of reflective (correct) answers (Frederick 2005).

Participants were then presented with a series of footbridge-type moral dilemmas and told to indicate for each case if they would choose the utilitarian option. There were six items, all taken from Koenigs et al. (2007), and selected on the basis of their emotional salience (*Baby*, *Allergy*, *Soldier*, *Submarine*, *Ecologists*, *Lifeboat*). Here is a sample item (see Appendix A for the rest):

> Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables.
>
> Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.
>
> Would you smother your child in order to save yourself and the other townspeople?

For each item, participants answered yes or no, and then indicated confidence in their choice on a 7-point scale. Their responses were then converted to a 14-point scale by combining the binary and confidence scales (-6.5 to -0.5 if they answered no, depending on confidence level, 0.5 to 6.5 if they answered yes). For each participant, responses for each item were then summed to yield a total score, with higher scores indicating a greater propensity to choose the utilitarian option.

2.2 Results

Correlational analysis revealed a clear relationship between utilitarian judgment and socially directed emotionality. Most striking were the negative correlations with measures of empathic concern ($r = -0.418$, $p = 0.002$, 2-tailed) and peer-reported empathy ($r = -0.369$, $p = 0.007$). Utilitarian judgment was also

positively correlated with callous affect ($r = 0.348$, $p = 0.011$), interpersonal manipulation ($r = 0.338$, $p = 0.013$), and aggressiveness ($r = 0.300$, $p = 0.029$). There was no significant correlation between utilitarian judgment and any of the other measures, including emotionality and emotion regulation (see Figure 6.3). Of the three peer measures, only the empathy measure, the measure most closely conceptually related to empathic concern, was significant. The r-squared values provide the best estimate of effect size for the different measures studied. Our claim is not that these measures might not correlate with utilitarian responses given a large enough sample, but rather that there is a much clearer and stronger relationship to measures of antisocial tendencies than to measures which Greene's account predicts as decisive. This is established both by the difference in magnitude of effect size and by the consistency of our findings with regard to various measures of antisocial tendencies as opposed to various measures of emotionality and emotion regulation.

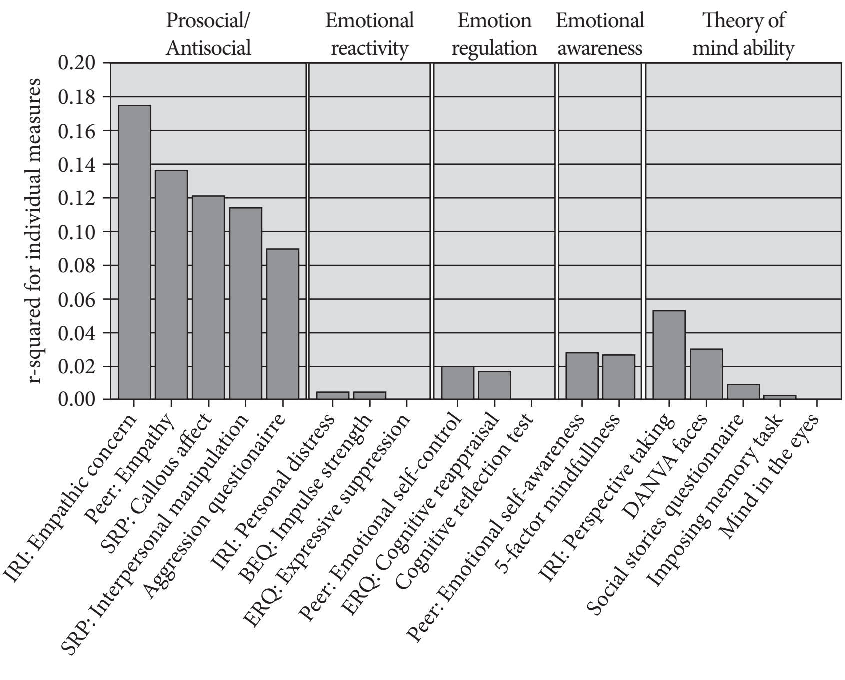


Figure 6.3 Correlations between utilitarian judgment and measures of emotional functioning and social intelligence (Study 1). Only the first group of measures (prosociality and antisociality) are significant. Y-axis shows r-squared (proportion of variance explained) for individual measures considered in isolation.

2.3 Discussion

The results of this first study replicate earlier findings of a negative relationship between utilitarian judgment and prosociality, and add a peer-report measure of empathy. The peer measure closes a small but significant gap in the literature. The patient studies do not offer clear specificity because those findings might reflect the influence of comorbid symptoms rather than antisocial tendencies. On the other hand, previous studies in neurotypicals have relied either on self-report measures (which measure self-concept) or proxy measures (e.g., genetic variation), and thus don't establish a direct link between utilitarian judgment and prosocial behavior. The peer measure does establish a clear link—individuals who are perceived by their peers to have greater empathy on the basis of their real-life behavior have a greater tendency to resist the utilitarian option. Further supporting the specificity of this effect, we observed no comparable correlations between utilitarian judgment and measures in any of the other categories investigated, including emotionality and emotional and cognitive self-regulation. This shows that, contrary to Greene's model, resistance to the utilitarian option is most strongly linked to prosocial emotion, rather than to intuitive emotional responding in general.

3 Study 2

Study 1 demonstrated an effect for the specificity of other-oriented and prosocial emotions, particularly compassion, and resistance to the utilitarian option in personal-harm scenarios. This experiment addresses specificity in a second way, namely by looking at the types of moral dilemmas in which reasoning is influenced by compassion. As noted earlier, the alternative to the utilitarian option in scenarios like those used in Study 1, which involve direct personal harm, is often labeled the deontological choice. It is unclear, however, whether the tendency to take the nonutilitarian option in footbridge-type scenarios coincides with a general tendency to conform with deontological principles. In other words, does compassion relate to deontological reasoning across the board, or only when personal harm is involved? To test this, we designed vignettes in which one option conformed to deontological rules (e.g., telling the truth, keeping your promises, and acting fairly), whereas

the other option violated a deontological rule (e.g., an abstract principle of justice or fairness) but promoted the greater good. In these vignettes, no act of direct personal harm to an individual was involved in either choice. It is important to note that these vignettes differ from the *prudential dilemma* used by Gleichgerrcht and Young (2013) insofar as the nondeontological choice is not personally self-serving.

3.1 Participants and procedure

97 participants (58 female, mean age 39.4 years) completed the study via a link from Amazon Mechanical Turk. After answering basic demographic questions (age, gender, level of education), they completed the full SRP and IRI scales, and the "No Meaning" scale (Kunzendorf et al. 1996). Finally, they answered yes/no questions for 18 vignettes. The vignettes consisted of the 6 items from Koenigs et al. (2007) used in Study 1 ("personal harm scenarios") plus 12 new vignettes ("deontological principle scenarios") (see Appendix B for details). In each vignette, participants were asked if they were willing to perform an action where answering yes corresponded to the utilitarian choice. The measure used was the number of "yes" responses to each vignette type. In addition there was an attention-check vignette included for the purposes of excluding participants who did not read each vignette carefully. A sentence in the middle of this vignette alerted participants to the fact that this was a test of attention and that they should click both yes and no to indicate they had read the entire text. Only participants who answered this question correctly were included in the participant count above.

3.2 Results

The correlation between "yes" responses to the personal harm and deontological principle scenarios was weak and nonsignificant ($r = 0.144$, $p = 0.15$), suggesting that different factors drive utilitarian judgments in the two types of scenario. As predicted from previous findings, the direct personal harm scenarios correlated significantly with empathic concern ($r = -0.321$, $p < 0.001$), callous affect ($r = 0.434$, $p < 0.001$), interpersonal manipulation ($r = 0.300$, $p < 0.005$), and No Meaning ($r = 0.249$, $p < 0.05$). In contrast,

none of these scales correlated with responses to the deontological principle scenarios ($|r| < 0.147$, $p > 0.15$). Looking at individual items, responses to all six of the personal-harm scenarios were significantly correlated with callous affect ($r > 0.231$, $p < 0.05$). There was no such correlation for any of the twelve deontological principle scenarios ($r < 0.153$, $p > 0.134$).

3.3 Discussion

The results of this study suggest that prosocial emotionality is not associated with a general tendency to judge in accordance with deontological principles. Rather, it relates specifically to thinking about situations in which an individual might be directly harmed in order to promote the aggregate welfare of the group. Likewise, antisocial traits like callous affect do not predict a resistance to deontological reasoning in general, but only in footbridge-type scenarios. These results replicate and extend those by Gleichgerrcht and Young (2013), who found that empathy was unrelated to participants' willingness to lie and cheat on their taxes. We use a broader range of scenarios, which are not confounded with the issue of choices being self-serving. In conjunction, these two studies support the claim that reduced empathic concern is linked with utilitarian responses to personal-harm scenarios specifically, rather than the tendency to endorse deontological reasoning generally.

4 Study 3

Studies 1 and 2 demonstrate a selective relationship between compassion and resistance to utilitarian reasoning in moral scenarios that involve using human lives as "means to an end" (e.g., sacrificing individuals in order to promote the general welfare). This experiment compares responses to footbridge-type dilemmas for humans and (nonhuman) animals, specifically dogs. Each scenario presents a choice between sacrificing an individual in order to save a group (the utilitarian option) versus sparing an individual and allowing a group to perish in the process (the deontological option). The original vignettes were modified so that the dog and human versions were as similar as possible without unduly compromising the plausibility of the vignette. In addition to

being assessed for prosocial and antisocial traits, participants were assessed for the tendency to differentiate between humans and dogs with respect to their morally relevant characteristics—the thought being that the presence of such a tendency might influence participants' responses to the two types of scenario.

4.1 Participants and procedure

281 participants (148 female, mean age 32.0 years) took part in the study using a link from Amazon Mechanical Turk. After answering basic demographic questions, participants completed a 20-item questionnaire that assessed the degree to which they differentiated between humans and dogs in terms of their value and mindedness, answering on a seven-point scale. Four questions directly compared dogs and humans (e.g., "Humans are worth more than dogs," "Humans have a higher level of consciousness than dogs"), and the remaining 16 items comprised 8 questions which were asked separately about dogs and humans (e.g., "Humans[/Dogs] understand the difference between right and wrong," "Humans[/Dogs] have a soul," "Humans[/Dogs] appreciate and reflect on their memories"). A composite score was calculated by taking the average of all 20 items, with the 8 questions specific to dogs reverse scored.

Participants then took the entire IRI scale, as well as the callous affect subscale of the SRP. Afterwards they answered yes or no to either the dog or the human version of 7 footbridge-type vignettes adapted from Koenigs et al. (2007), modified so that the dog and human versions were as similar as possible, with participants randomly assigned to either the dog or the human version. An attention-check vignette was used to exclude participants who failed to read all texts in their entirety.

4.2 Results

Participants in the dog condition ($n = 135$) chose the utilitarian option (mean = 4.23, s.d. = 1.50) more frequently than participants in the human condition ($n = 146$, mean = 3.04, s.d. = 1.85). The difference between conditions was highly significant ($t[274.1] = 5.94$, $p < 0.001$, equal variances

not assumed). There were no significant differences in demographic or personality measures between the two groups.

For the human scenarios, utilitarian choices were significantly correlated with empathic concern ($r = -0.356$, $p < 0.001$) and callous affect ($r = 0.460$, $p < 0.001$). For the dog scenarios, utilitarian choices were not correlated with empathic concern ($r = -0.094$, $p = 0.278$), and only marginally significantly correlated with callous affect ($r = 0.152$, $p = 0.079$).

We next divided the data to look separately at individuals who clearly differentiate humans and dogs, and those who regard humans and dogs as being more similar with respect to mindedness and value. Looking at the entire data set (across both conditions) there was no correlation between scores on the human-dog differentiation scale and empathic concern, callous affect, or number of utilitarian choices ($|r| < 0.090$, $p > 0.13$).

There was a trend for individuals who differentiate less between humans and dogs to be less willing to choose the utilitarian option both for humans (high differentiators mean = 3.26, s.d. = 1.836; low differentiators mean = 2.80, s.d. = 1.854) and for dogs (high differentiators mean = 4.36, s.d. = 1.515; low differentiators mean = 4.10, s.d. = 1.478). However, this difference was not significant for either condition ($t < 1.516$, $p > 0.132$). The more significant finding related to how compassion modulated responses in the two cases.

The patterns of correlation between empathic concern, callous affect, and utilitarian choices are summarized in Tables 6.2 and 6.3. Notably, for

Table 6.2 Correlations between utilitarian judgment (number of yes responses) and empathic concern (Study 3)

	Differentiate less	**Differentiate more**
Human	−0.359** ($n = 75$)	−0.358** ($n = 70$)
Dog	−0.086 ($n = 62$)	−0.146 ($n = 68$)

* indicates $p < 0.05$; ** indicates $p < 0.01$.

Table 6.3 Correlations between utilitarian judgment (number of yes responses) and callous affect (Study 3)

	Differentiate less	**Differentiate more**
Human	0.460** ($n = 75$)	0.438** ($n = 70$)
Dog	0.266* ($n = 62$)	0.076 ($n = 68$)

* indicates $p < 0.05$; ** indicates $p < 0.01$.

those who differentiate more clearly between humans and dogs (median split), correlations between compassion and number of utilitarian responses remained approximately the same for humans, but were no longer significant for dogs. On the other hand, for those who differentiate less, a significant correlation with callous affect was still present for dogs, but of lower magnitude, accounting for only 25 percent of the variance accounted for in the human case.

4.3 Discussion

People are generally more willing to choose the utilitarian option in the case of dogs than in the case of humans.[6] Further, compassion is associated with resistance to the utilitarian option in the case of humans, but not in the case of dogs. This suggests that the key issue with respect to the influence of compassion is whether the action in question specifically involves harming a human being, as opposed to harming a sentient creature of one sort or another. The specificity here, however, appears to depend upon the extent to which one thinks of human beings as "special" in terms of mindedness and moral status. This result echoes one of the central themes of the literature on dehumanization: the idea that thinking of an individual as lacking essentially human psychological features (a human psychological "essence") goes along with regarding that individual as less worthy of protection from intentional harm (Haslam 2006; Smith 2011; Leyens et al. 2001).

5 Study 4

Studies 1, 2, and 3 demonstrate a consistent negative relationship between compassion and utilitarian reasoning, at least in scenarios where participants are asked whether they themselves, with a single decisive act, would sacrifice another human being. This study examines a different type of moral dilemma, namely, scenarios involving "passive euthanasia," in which a patient is allowed to die by discontinuing his or her medical treatment. Passive euthanasia is distinguished from euthanasia of the "active" variety, in which the patient is intentionally killed by direct means (see Study 5, below). This practice has not been legally classified as euthanasia in the United States for decades. It is legally

classified as withholding or withdrawing medical treatment—treatment that patients or their guardians are, in most cases, legally entitled to refuse.

The present study examines whether an agent's level of compassion predicts their stance toward passive euthanasia, especially in cases where an individual is suffering and likely to die soon. In such cases, the utilitarian (welfare-maximizing) perspective favors euthanasia, whereas the deontological (persons-as-ends) perspective opposes it. In contrast to the previous three studies (and Study 5, below), in which participants were asked what they themselves would do in a given situation, participants in this study were asked to judge the moral acceptability of an action directed by one stranger (the patient's family) toward another (the patient). This is an important distinction, because resistance to euthanasia in this case cannot be explained by the participant's immediate desire to avoid personally participating in an act of harming.

5.1 Participants and procedure

The sample is the same as in Study 1, but with more participants ($n = 70$, rather than $n = 53$), since more people filled out this questionnaire than the footbridge questionnaire. Of the 53 participants who did both, the correlation between the utilitarian choices (i.e., pushing the fat man and allowing the patient to die) was strong and highly significant ($r = 0.515$, $p < 0.001$).

Measures of prosocial and antisocial emotionality, emotional expressiveness, emotion regulation, and social intelligence were completed as in Study 1 (see Section 2.1, above). The euthanasia survey was filled out online between laboratory testing sessions. There were two sets of questions, one involving infants and the other involving the elderly. Prior to the questions about the infant, the survey reads: "Do you believe that parents, consulting with their pediatrician, should be permitted to discontinue medical treatments that may preserve the life of a week-old newborn, if. . . ." This was followed by eight specific scenarios, such as "the newborn has an able brain and body, but has a condition that will cause constant pain for the rest of his or her life." Prior to the questions about the elderly person, the survey reads: "A 70-year-old person, who has not previously expressed an opinion toward whether s/he would want to be kept alive, has fallen into a coma. Should the person's relatives

be permitted, in consultation with a doctor, to discontinue medical treatments that may preserve the person's life, if. . . ." This was followed by seven specific scenarios, such as "the person may awake from the coma, but have severe and disabling disfigurement of the face, arms and legs."

5.2 Results

The pattern observed was very similar to that seen in Study 1 for utilitarian judgments—that is, all and only prosocial and antisocial measures were significant (empathic concern, $r = 0.579$, $p < 0.001$; callous affect, $r = 0.333$, $p = 0.015$; interpersonal manipulation, $r = 0.308$, $p = 0.025$; aggressivity, $r = 0.313$, $p = 0.022$)—with two exceptions. First, the peer-reported empathy correlation ($r = -0.236$, $p = 0.089$) fell just short of significance, though it can be considered significant at an alpha of 0.05 insofar as we can think of the hypothesis as one-tailed. Second, proeuthanasia responses were significantly correlated with personal distress ($r = 0.372$, $p = 0.006$).

Reponses to all of the individual questions were significantly correlated with both empathic concern and callous affect, with the exception of the two items that involved the person never regaining consciousness (but even these items trended in the same direction).

5.3 Discussion

It appears that prosocial emotionality goes hand in hand with a tendency to reject passive euthanasia, even in cases where the individual concerned would be in pain for the rest of her or his life and would never recover even minimal cognitive or communicative function. This is surprising, in that it seems natural to think of euthanasia in such cases as being the compassionate choice. It turns out that the more compassionate you are, the more inclined you will be to keep a person alive, even if the life you preserve is full of misery and devoid of future promise.

In terms of theoretical implications, it is noteworthy that personal distress is positively associated with endorsement of passive euthanasia—even in cases where the individual whose life is at stake is a stranger, and personal distress at the prospect of his or her death seems less likely to be a factor in

judgment than it would be otherwise. This runs counter to the idea, central to Greene's model of moral judgment, that raw emotional responding (of which personal distress is paradigmatic) drives resistance to the utilitarian option.

6 Study 5

Here we examined cases of active euthanasia involving both humans and dogs, as opposed to cases of passive euthanasia involving humans only, as in the previous study. Active euthanasia (sometimes colloquially described as "mercy killing") involves intentionally causing a person's death in order to relieve his or her suffering. The primary goal of the study was to determine whether attitudes toward active euthanasia are correlated with socially directed emotion traits. Of secondary interest was whether such correlations, if observed, would differ in the case of humans and the case of dogs, possibly as a function of the degree to which people distinguish between humans and dogs with respect to morally relevant features of their psychology.

6.1 Participants and procedure

There were 86 participants (59 female, mean age 37.1 years), directed to the study by a link from Amazon Mechanical Turk. After answering basic demographic questions (age, gender, education), participants filled out the full SRP, the full IRI, and the human-dog differentiation scale used in Study 3. They then answered yes or no to the probe question of one of the following two vignettes, to which they were randomly assigned:

> *Dog condition:* You are hiking along a trail, many miles from the nearest road. You hear a strange noise, and decide to investigate. You find a dog a few steps off the path has been attacked by a large predator and is bleeding to death. The dog is whimpering and whining, and is obviously in overwhelming pain. You have seen injuries like this before, and you know that, left alone, the dog will endure a slow and painful death. There is nothing you can do to save the dog's life. However, you carry a gun with you whenever you go hiking.
>
> Would you shoot the dog to put him out of his misery?

Human condition: You're hiking along a trail, many miles from the nearest road. You hear a strange noise, and decide to investigate. You find a man a few steps off the path. He has been attacked by a large predator and is bleeding to death. He is in obvious pain, but is unable to speak apart from crying. You have seen injuries like this before, and you know that, left alone, the man will endure a slow and painful death. There is nothing you can do to save the man's life. However, you carry a gun with you whenever you go hiking.

Would you shoot the man to put him out of his misery?

As before, an attention-check vignette was used to exclude participants who did not fully read the text.

6.2 Results

There were 86 participants (59 female, mean age 37.1 years), randomly assigned to the two conditions (dog scenarios $n = 46$, human scenarios $n = 40$). Neither demographic nor personality variables differed significantly by condition.

There was a marked tendency for people to be more willing to kill a dog in order to put it out of its misery (mean = 0.63, s.d. = 0.488, expressed as a proportion from 0 to 1), than were willing to kill a human (mean = 0.10, s.d. = 0.304). The difference was highly significant (Fisher exact test of 2×2 contingency table, $p < 0.0001$). For the human condition, there was a significant correlation between proportion of kill responses and both empathic concern ($r = -0.479$, $p < 0.005$) and callous affect ($r = 0.555$, $p < 0.001$). For the dog condition, there were no significant correlations between proportion of kill responses and empathy measures (empathic concern $r = -0.204$, $p = 0.175$; callous affect $r = 0.187$, $p = 0.213$).

We next divided the data to look separately at individuals who clearly differentiate between dogs and humans in terms of mindedness and moral status, on the one hand, and individuals who regard dogs and humans as more equal in status. Looking at the entire data set (across both conditions) there was no correlation between scores on the human-dog differentiation scale and empathic concern, callous affect, or number of proeuthanasia responses ($|r| < 0.057$, $p > 0.602$).

The patterns of correlation that emerged between empathic concern and proeuthanasia responses are summarized in Tables 6.4 and 6.5. What is striking here is that the correlation between compassion and resistance to

Table 6.4 Correlations between approval of active euthanasia (number of yes responses) and empathic concern (Study 5)

	Differentiate less	**Differentiate more**
Human	−0.479** ($n = 17$)	−0.669** ($n = 23$)
Dog	−0.515** ($n = 25$)	0.508* ($n = 21$)

* indicates $p < 0.05$; ** indicates $p < 0.01$.

Table 6.5 Correlations between approval of active euthanasia (number of yes responses) and callous affect (Study 5)

	Differentiate less	**Differentiate more**
Human	0.555** ($n = 17$)	0.716** ($n = 23$)
Dog	0.468* ($n = 25$)	−0.258 ($n = 21$)

* indicates $p < 0.05$; ** indicates $p < 0.01$.

active euthanasia reverses for animals that are clearly regarded as inferior to humans in terms of mental status.

6.3 Discussion

In general, participants were much more willing to commit active euthanasia in the case of dogs than in the case of humans. For those who regard dogs as inferior in mindedness and value to humans, compassion is associated with greater approval of euthanasia in the case of dogs. For those who regard dogs as more or less on a par with humans, however, compassion predicts resistance to euthanasia in the case of both dogs and humans. These findings, like the findings of the previous study, run counter to the intuitively plausible idea that the more compassionate you are, the more inclined you will be to put a suffering being, human or otherwise, out of its misery. Here as elsewhere, compassion predicts resistance to means-end moral reasoning of the utilitarian variety, at least where humans (and animals perceived as relevantly human-like) are concerned.[7]

7 General discussion

7.1 Summary

The broad upshot of our behavioral investigations is this: Prosocial emotion in particular, rather than emotional responsivity in general, is associated

with resistance to instrumental moral decision-making of the kind endorsed by utilitarians, at least in cases where the action contemplated involves directly harming either a human being or a being perceived as having relevantly similar status. This means that, up to a point, the deontological perspective on morality is affiliated with specifically prosocial emotionality in a way that the utilitarian perspective is not. This comes as something of a surprise, given that it seems natural to think of the utilitarian stance toward euthanasia, at least in certain extreme cases (e.g., terminal cancer patients in severe pain that cannot be controlled by medication), as more "humane" than the deontological alternative. What it suggests, however, is that the typical person's aversion to high-conflict utilitarian choices is driven by an intuitive sense of the sanctity of human life, not just the value of human life in the aggregate. This is a far cry from the "emotional alarm" model of deontological judgment in Greene's dual-process account, according to which raw emotional responding drives resistance to the utilitarian option (recent psychopharmacological evidence also tells against Greene's view; see Terbeck et al. 2013). Indeed, as we saw in Study 4, where personal distress was associated with endorsement of passive euthanasia, the "emotional alarm" tends if anything to have the opposite effect—it encourages utilitarian responding. Whereas Gleichgerrcht and Young (2013) found that personal distress was unrelated to utilitarian responses to both personal and impersonal scenarios, our results suggest that, depending on the details of the scenario, this sense of "emotional alarm" can have different effects on moral decision-making.

In this connection it is important to note a point of contrast between Greene's dual-process model and ours. In Greene's model, the psychological processes driving deontological judgment are situated at the level of primitive, reflexive, "lower" emotion. As noted earlier, Greene has likened this emotional response to the workings of our "inner chimp" (Radiolab 2007). In our view, by contrast, the emotional responding that pushes people in a deontological direction is quite different. First, the emotional push derives from a highly refined aspect of human nature—a psychological motivation that has traditionally been thought to distinguish us from other animals (but see De Waal 2009, for a softening of this perspective). In colloquial terms, it is not our "inner chimp" but our sense of humanity that motivates resistance to utilitarian thinking. Second, it involves a high degree of representational sophistication and cognitive flexibility, which are absent from Greene's picture

of how deontological judgment arises. This seems clear from the observation (in Studies 3 and 5) that the tendency to choose the deontological option in hypothetical scenarios is sensitive to differences in the mental status of the characters involved. If deontological judgment were simply a matter of raw emotional responding, as Greene has it, it seems unlikely that it would show this kind of sensitivity. Instead, the pattern of judgments observed in our studies suggests the possibility that deontological judgment reflects intuitive appreciation of an abstract moral principle prohibiting intentional harm to human or sufficiently human-like beings, that is, something akin to Kant's "principle of humanity" (the idea that rational, autonomous beings should be treated as ends-in-themselves, not merely as means to an end). Prosocial emotions like compassion, then, may be essentially bound up with this appreciation.

7.2 Dehumanization and experiential understanding

If the perception of mindedness is linked to an intuitive appreciation similar to Kant's "principle of humanity," then this psychological mechanism would help to explain why members of dehumanized out-groups are so often denied justice, or even subject to intentional harm. Because such individuals are regarded as less than fully human (e.g., lacking a proper "human essence"), they are seen as falling outside the scope of Kant's principle, and their plight does not elicit the same response afforded to the fully human. Attempting to capture what is missing in the perception of a dehumanized out-group, Gaita writes about the legal treatment of Aboriginals in Australia as follows:

> We love, but they 'love,' we grieve but they 'grieve' . . . we may be dispossessed but they are 'dispossessed.' That is why, as Justice Brennan said, racists are able 'utterly to disregard' the suffering of their victims. If they are to see the evil they do, they must first find it intelligible that their victims have inner lives of the kind which enable the wrongs they suffer to go deep. (Gaita 1999, p. 78)

What Gaita eloquently evokes here is a curiously ineffable missing essence—an essence we tend to perceive as an integral aspect of the psychological states of in-group members but not of dehumanized out-groups. Our recent findings suggest that humanizing depictions evoke the neural signature of the phenomenal

stance, whereas animalistic dehumanizing depictions, which involve a socially distanced representation of the emotional states of others, evoke the intentional stance (Jack et al. 2013b).

Gaita's remarks correspond well with our original statement of the phenomenal stance (Robbins and Jack 2006), a mode that we claim is essential for us to appreciate the ineffable phenomenal characteristics of mental states. Our claim was, and still is, that we need the phenomenal stance to understand the cognitive origins of the "hard problem" of consciousness. Consider, for example, the philosophical zombie thought experiment—an "intuition pump" used to highlight the hard problem. In this thought experiment, by hypothesis, the anomalous individual outwardly behaves as if he or she experiences and perceives the world as we do. Yet, by hypothesis, they have no experiential world. They "love" but they do not love. While dehumanizing depictions have the surreptitious effect of undermining our sense of intersubjective understanding, the zombie thought experiment pushes explicitly on this idea.

7.3 The language of morality

The most challenging question raised by our findings relates to the case of euthanasia. How might we best describe, in ordinary language, the moral response we have to euthanasia cases involving humans, but not to euthanasia cases involving subhuman animals? Our findings indicate that a virtuous prosocial emotion so often connected with the motivation to ameliorate another person's suffering—compassion—does not always manifest itself as such. For the case of euthanasia, high levels of compassion may actually counteract the concern to alleviate human suffering. The counterintuitive nature of this finding makes the description of this moral response a surprisingly vexing question. At first glance, it might seem tempting to describe the no-kill option as the "compassionate" or "humane" response. However, this move would lead to misunderstanding. Both these terms have been co-opted to refer to the positive motivation behind acts of euthanasia, which must be distinct from the motivation we document here that causes resistance to euthanasia.[8] Surprisingly, it appears the ordinary language of humanism is at a loss to adequately describe the motivation revealed by our empirical findings. Yet more surprising, it appears that religious (or supernatural) language does a

better job. The best language we have been able to find to describe the motivation we document here is that it reflects a desire to preserve the "sanctity of human life." Alternatively, our euthanasia findings might be interpreted as suggesting that compassionate people weigh "preserving the soul" of the individual more heavily than preventing their suffering—that they regard persons as somehow greater than the sum of their psychological parts.[9]

We may be wrong to think that religious or supernatural language best captures this motivation. Perhaps we are simply lacking in imagination, failing to find the right words, or perhaps future studies will give this motivation a different spin.[10] We welcome alternative attempts to describe this moral motivation in more naturalistically acceptable terms. In the meantime, we note that other secular thinkers (e.g., moral philosophers) who have discussed the ethics of harm have also felt the need to appeal to notions of sanctity and the soul (McMahan 2002). This tendency is sufficiently pervasive that it is sometimes decried in discussions of adult resistance to science (Bloom and Weisberg 2007). Yet, we might want to be careful about legislating against the use of this type of language in moral discourse. Extrapolating from our findings, it is plausible to suppose that such a move would alienate many of the most prosocial members of our society, by being dismissive of their intuitions. It is also likely to bias moral discourse away from deontological considerations and toward utilitarianism. The notion that moral discourse should only employ scientifically respectable terms assumes that the language of morality and the language of science can be merged. We might at least consider an alternative view, namely, that the language of morality is at odds with the language of science. When we pause to consider the neuroscientific evidence, this view would certainly seem well motivated: scientific thought and moral thought rely on brain networks which are not only distinct, but which actively suppress each other (Figure 6.2). Hence, rather than assume these different languages and perspectives can be brought seamlessly together, we might entertain the notion that religious and supernatural language continues to flourish not in spite of the fact it differentiates itself from scientific language, but rather because it differentiates itself. That is, the intrinsically nonempirical nature of religious and supernatural language may actually facilitate the emergence and communication of ideas that derive from this alternative, and opposed, mode of understanding. This would be consistent with the schism in understanding

predicted by our theoretical model, illustrated in Figure 6.1 (Robbins and Jack 2006; Jack and Robbins 2012; Jack forthcoming). We hypothesized the tension between the phenomenal stance and the physical stance precisely to account for the tendency to adopt the supernatural belief of mind-body dualism. The idea that this can be generalized to other supernatural beliefs is supported by recent findings of ours that there is a robust and highly replicable tendency for individuals with greater empathic concern to believe more in God, whereas those who score higher on analytic thinking believe less (Jack et al., under review). The connection between prosocial versus analytic cognition and religious belief has also be found on Twitter, where religious individuals use more social and emotional vocabulary than atheists, who use a more analytic vocabulary in their "tweets" (Ritter et al. 2013). These data point to the notion that religious beliefs are associated with social connection and prosocial emotions in a way that nonreligious sentiments are not. Putting these pieces together, we believe there is good reason to seriously consider the hypothesis that key aspects of moral reason lie beyond the grasp of analytic-empirical thinking but can be quite readily captured by religious language.

7.4 What normative significance for science?

What significance does this work have for normative projects? Clearly, our model differs from Greene's not only in its origins and on a number of important points, but also in terms of our normative disposition. Greene is a utilitarian, whereas we think utilitarianism misses some essential aspects of moral understanding; and we think this was an insight that Kant, among others (e.g., Bernard Williams; see Williams 1973), appreciated and did his best to build upon. It is not that we think the desire to maximize utility plays no role in moral decision-making. Other things being equal, surely the best action is the one that results in the best overall outcome. The question is how to weigh this goal against other moral considerations, which are typically harder to capture and define. We have no interest in defending wholesale Kant's conception of how to do this and what it involves. Kant is an important figure to us primarily because he famously articulated a position that resists utilitarianism, and we think there is wisdom to be found in that perspective, regardless of whether he got the details right.

Of course, there is no external standard of correctness against which our normative views and Greene's can be measured. As has been previously argued in the context of Greene's work, science is not going to deliver such a standard (Berker 2009). However, it does not follow that this scientific work is insignificant for normative projects. On the contrary, we think there is considerable value to examining how closely these two competing characterizations of the processes leading to moral judgment fit the neuroscientific and psychological data. Why? Well, let us suppose that Greene's account is not merely scientifically inaccurate, but that it also evokes a kind of rhetoric that inclines us to dismiss the Kantian perspective and adopt Greene's instead. Whether it was his intention or not, Greene has associated resistance to utilitarianism with undesirable ways of thinking—with a base emotionality that can either be construed as dumb—"the automatic setting on a camera" (Greene 2010)—or as reflecting our animal nature —"our inner chimp" (Radiolab 2007). Such emotive descriptions would be understandable indulgences if those aspects of his model looked robust, but they become problematic in light of the issues raised in this chapter. Contrary to Greene, it appears that resistance to utilitarianism is not automatic, emotion driven, or primitive; instead it derives from our highly evolved and reflective capacities for social and emotional insight. Resistance to utilitarianism is associated with prosocial behavior, whereas utilitarianism is associated with what may be the least desirable personality type of all—that of the psychopath.[11] These observations are not, of course, sufficient to determine which perspective is normatively superior. But they do suggest that one side has been presenting a misleading picture. In the long-running battle to win converts to one or the other of these ethical perspectives, it is not so much the scientific details in themselves that matter, but what those details tell us about the desirability of different ethical systems. It seems safe to say that pragmatic concerns such as these are significant from a normative point of view.

Acknowledgments

This work was supported in part by funding to AIJ from the Leonard Krieger Fund and the University Hospitals Case Medical Center Spitz Brain

Health fund. The authors would like to thank Abigail Dawson, Joshua Gordon, Regina Leckie, and Megan Norr for assistance with experimental design and data collection; and Jeremy Bendik-Keymer, Edouard Machery, Garrett Marks-Wilt, and Stuart Youngner for comments on an earlier draft.

Author contributions

AIJ oversaw the entire project, conceived and designed the experiments, assisted with the analysis, prepared figures, and contributed to the writing, in particular theoretical sections. PR reanalyzed the data and was the lead writer. JPF assisted with experimental design, collected and analyzed the data, and contributed to writing the methods. CDM conceived and designed the stimuli used in Study 2. All authors assisted in editing the manuscript.

Appendix A: Direct personal harm vignettes (Studies 1 and 2)

Allergy. You are a waiter. You overhear one of your customers say that he is about to go to jail and that in his last forty-eight hours of freedom he plans to infect as many people as possible with HIV. You know him well enough to know that he is telling the truth and that he has access to many potential victims.

You happen to know that he has a very strong allergy to poppy seeds. If he eats even one he will go into convulsions and have to be hospitalized for at least forty-eight hours.

Would you cause this man to have a serious allergy attack in order to prevent him from spreading HIV?

Ecologists. You are part of a group of ecologists who live in a remote stretch of jungle. The entire group, which includes eight children, has been taken hostage by a group of paramilitary terrorists. One of the terrorists takes a liking to you. He informs you that his leader intends to kill you and the rest of the hostages the following morning.

He is willing to help you and the children escape, but as an act of good faith he wants you to kill one of your fellow hostages whom he does not like. If you refuse his offer all the hostages including the children and yourself will die. If you accept his offer then the others will die in the morning but you and the eight children will escape.

Would you kill one of your fellow hostages in order to escape from the terrorists and save the lives of the eight children?

Lifeboat. You are on a cruise ship when there is a fire on board, and the ship has to be abandoned.

The lifeboats are carrying many more people than they were designed to carry. The lifeboat you're in is sitting dangerously low in the water—a few inches lower and it will sink.

The seas start to get rough, and the boat begins to fill with water. If nothing is done it will sink before the rescue boats arrive and everyone on board will die. However, there is an injured person who will not survive in any case. If you throw that person overboard the boat will stay afloat and the remaining passengers will be saved.

Would you throw this person overboard in order to save the lives of the remaining passengers?

Soldier. You are the leader of a small group of soldiers. You are on your way back from a completed mission deep in enemy territory when one of your men has stepped in a trap that has been set by the enemy and is badly injured. The trap is connected to a radio device that by now has alerted the enemy to your presence. They will soon be on their way.

If the enemy finds your injured man they will torture him and kill him. He begs you not to leave him behind, but if you try to take him with you your entire group will be captured. The only way to prevent this injured soldier from being tortured is to shoot him yourself.

Would you shoot this soldier in order to prevent him from being tortured by the enemy?

Submarine. You are the captain of a military submarine traveling underneath a large iceberg. An onboard explosion has caused you to lose most of your

oxygen supply and has injured one of your crew who is quickly losing blood. The injured crew member is going to die from his wounds no matter what happens.

The remaining oxygen is not sufficient for the entire crew to make it to the surface. The only way to save the other crew members is to shoot dead the injured crew member so that there will be just enough oxygen for the rest of the crew to survive.

Would you kill the fatally injured crew member in order to save the lives of the remaining crew members?

Appendix B: Deontological principle vignettes (Study 2)

Gun. Several months ago your neighbor, Jim, asked you if he could temporarily keep his gun in the safe at your house, for safety reasons, while his young niece and nephew stayed with him. The children came and went and your neighbor never got around to reclaiming his gun.

Today your neighbor comes and bangs on your door. You hear him muttering and cursing. When you answer the door he is obviously angry—his face flush and his fists clenched. He bypasses the usual pleasantries and gets straight to the point: he has come to get his gun back. Although you have never known him to be a violent person, you have also never seen him this angry. You are afraid he may hurt someone if you give him his gun.

Would you lie to Jim and tell him that you have lost the key to your safe?

Professor. You are a college professor teaching a class of 150 students. One student, Peter, comes to your office and begs you to change his final grade from a D to a B. Peter is smart, hardworking, and very responsible but struggled in the class because he was working two jobs to help support his out-of-work single mother and two younger sisters.

Peter has been offered a very lucrative job opportunity on the condition that he graduates this semester. If he does not receive at least a B in your class, he will be denied the job. It is unlikely that he will have such a lucrative offer any time in the future. No one else would know if you changed the grade.

Would you change Peter's grade to a B in order to help him and his family?

Wood Chipper. You are a judge presiding over a civil court case. The plaintiff, Mr Singer, lost his arm to a wood chipper. Singer has no medical insurance and has not been able to find gainful employment due to his disability. He is suing the manufacturer, ACME. ACME has a history of blatant disregard for consumer safety, but so far their top-notch lawyers have managed to protect them from liability. As it turns out, this time ACME was not at fault.

If you decide in favor of the plaintiff, ACME will pay more attention to consumer safety in the future. Furthermore, ACME would not protest the judgment, since it would cost more in legal fees to appeal than simply to pay the damages.

Would you decide in favor of Singer, even though ACME was not at fault?

Hiring. You are a manager at a high-tech company. Along with the regular workers there are two interns, Kate and Helen. Both are intelligent and hard-working recent engineering graduates looking for permanent jobs. You were impressed by them and wrote highly favorable letters of recommendation for both.

Kate and Helen are now finalists for the same unique and highly paid position. Someone from the organization contacts you and asks you which one is more qualified. Kate is much friendlier than Helen who can be abrasive and short tempered. Furthermore, Kate plans to donate a significant portion of her salary to a children's hospital. For these reasons, you would prefer to recommend Kate for the job. But the job does not involve working with others and Helen is actually the more qualified applicant.

Would you say that Kate is more qualified to help her get the job over Helen?

Bonus. You are a mid-level executive at a large company. Due to an unusually profitable year, corporate headquarters has allocated more money than usual for holiday bonuses. You have $20,000 to distribute to the two branch managers who report to you. The purpose of the bonus is to reward productivity. Both division managers work equally hard and are equally deserving.

One of the division managers, Mr Norcross, works in your office. His house was recently destroyed in a flood and his mother needs an operation. The other manager, Mr Kamm, has been luckier than Norcross and lives quite comfortably on his salary. You would like to give Norcross a $15,000 bonus

and give $5,000 to Kamm. Norcross and Kamm have never met, so neither will know what the other received.

Would you give Norcross the bigger bonus to help him out?

Painting. You are at the hospital visiting your Aunt Doris, who is in a coma. She unexpectedly wakes up and asks you to perform a dying wish for her. She informs you that in her will she has bequeathed to you her collection of landscape paintings, which are nice but not especially valuable. Now, she tells you that she wants you to give a particular painting, the desert landscape in the foyer, to her country club. You promise to do so. Doris falls back into her coma, and dies just hours later.

At the wake you come across Gloria, Doris's faithful and kind housekeeper of 20 years. She is in the foyer admiring the landscape painting. She tells you how she has always loved this particular painting, and will miss seeing it. You realize that Gloria would get much more pleasure from the painting than the members of Doris's country club.

Would you give the painting to Gloria in spite of your promise?

Subway. You are just outside the subway station on your way to work when you realize that you do not have any money for the train. It would take 45 minutes to walk to work. If you walked, you would be late and your boss and coworkers would be very irritated and disappointed. Furthermore, it is unseasonably warm, and so you would arrive at work hot, sweaty, and tired.

You notice that the gate for wheelchair access to the subway platform has been left open. There is no one around. You realize that you could easily sneak through the gate unnoticed and ride for free.

Would you sneak into the subway station to avoid having to walk to work?

Bone Marrow. You are the Chief of Medicine at a hospital. A patient named Mr Nozick was recently diagnosed with lymphoma, but before treatment began the tumors vanished and he was cancer-free. His doctors believe that Nozick's body has a unique ability to fight cancer and that examination of his bone marrow could help countless future cancer victims.

You ask Nozick if he would donate some of his marrow for research. Despite your promises of compensation and pleas on behalf of those who might benefit, he refuses. Although the procedure will be only mildly painful, Nozick has an

unusually strong aversion to pain. Another doctor suggests telling Nozick that he needs an appendectomy, and, when he is under anesthesia, the surgeons will remove some of his marrow without telling him.

Would you lie to Nozick and allow his doctors to take some of his marrow without his consent?

Transplant. You are the chief of surgery at a hospital. Two patients, Mr Greene and Mr Brown, both need a new heart very soon, but only one is available. It is your job to examine their records and decide which patient the surgical team will operate on. Brown is a 78-year-old recluse with no family. He is first on the waiting list. Greene is a 38-year-old widower raising two young daughters. Without the heart, Greene's daughters will be orphaned and sent to foster homes.

To ensure fairness, hospital policy requires that organs go to the patient at the top of the list. However, a patient could be denied if surgery is too risky. Brown's surgery would not be any riskier than Greene's. Nevertheless, no one would question your expert opinion if you were to claim otherwise and give the heart to Greene.

Would you falsely claim that surgery is too risky for Brown in order to give the heart to Greene?

District Attorney. You are the district attorney prosecuting a violent career criminal, Mr Smith, for deliberately starting a fire in which an elderly couple died. Two years earlier Smith murdered a nun, but he was not convicted due to a technicality. If convicted, Smith will spend the rest of his life in prison. If not convicted, Smith will go on to commit other serious crimes.

You have evidence that proves Smith is not guilty. The fire was actually started accidentally by one of the elderly residents who died in the blaze. As DA, it is your legal obligation to turn this evidence over to the defense. However, if you suppress this evidence, no one will ever find out and Smith will be convicted.

Would you suppress the evidence in order to ensure that Smith goes to prison?

Pasta. You have recently started making friends with a coworker. She invites you to her house for dinner. When you arrive you can see from the kitchen

that she has obviously put significant effort into preparing the meal. When you sit down to eat, however, you do not like the food at all. She made pasta puttanesca, even though you told her more than once that you do not like olives. Still, you manage to eat it anyway, out of politeness and focus on the enjoyable conversation.

About half way through the meal she asks you, with eager anticipation, whether you like the food. It is obvious that she would be disappointed by anything less than raving praise.

Would you lie and tell her the food is delicious in order to spare her feelings?

Restaurant. You were recently hired at an entry-level position at a finance company. One evening your boss invites you to dinner with several other executives. Seeing this as a great opportunity to advance your career, you accept the invitation.

The restaurant is quite extravagant and your meal costs significantly more than you were expecting. Much to your horror, you find that you do not have nearly enough cash on you, and the restaurant does not take credit cards. When the check comes, everyone else pitches in. They then give the money and the bill to you. There is already enough to cover the entire bill and provide a generous tip. No one would notice if you did not add any money. To pay for your dinner you would have to ask your boss to lend you money, which would make you both very uncomfortable.

Would you use the other diners' money to pay for your meal?

Notes

1 Russell quotes from *Education and the Good Life*, by Bertrand Russell, copyright © 1926 by Bertrand Russell. Reproduced by kind permission of The Bertrand Russell Peace Foundation, Ltd and Taylor and Francis Group.
Zajonc quotes from *Emotions*, edited by D. T. Gilbert and S. T. Fiske, copyright © 1998 by D. T. Gilbert and S. T. Fiske. Reproduced by kind permission of Dr. Hazel Markus.

2 Hence true moral engagement may be distinguished from moralizing, which does not seem to involve this.

3 Greene uses the term "cognitive" in scare quotes to characterize these processes. We use the term "analytic" to emphasize the contrast with a more synthetic style of thinking involved in empathetic understanding. These are differences in emphasis. There is little dispute concerning the broad characterization of these processes.

4 In line with this, it is worth noting that if an individual were to respond to the scenarios in the manner that Greene endorses as rational and preferable then their actions would, in nearly every case, make them guilty of murder or manslaughter. Hence, Greene's view of good moral reasoning diverges both from what most educated Westerners do when faced with ethical dilemmas, and is at variance with all or nearly all of the world's established civilian legal systems.

5 A reviewer of this chapter concluded that the implications of these findings are that "compassion appears to be a fairly blunt emotion . . . being . . . unable to take into account suffering." This is a misunderstanding. The measures used directly measure the individual's sensitivity to the suffering of others (as can be ascertained by reviewing the items, and a very large psychological literature). This is what makes it so surprising that these measures nonetheless predict even greater sensitivity to concerns about sanctity of life. It does not follow from the fact that something else trumps suffering that there is no sensitivity to it.

6 Note that this finding is consistent with the idea that judgments about the moral standing ("moral patiency") of individuals are distinctively tied to our perception of those individuals' capacity for phenomenal experience (Robbins and Jack 2006; Jack and Robbins 2012). Phenomenal experience covers a wide range of psychological states: not just basic affective experiences such as pleasure and pain (the primary concerns for the utilitarian), but also higher, social emotions (e.g., pride, guilt, and embarrassment) that we tend to think of as uniquely human.

7 Like the results of Study 3, these findings are compatible with the claim that perceptions of moral standing covary with perceptions of experiential mindedness (see Footnote 6, above). The fact that compassion does not typically predict resistance to the utilitarian choice where dogs are concerned, for example, may reflect the fact that people typically think of the experiential mental life of dogs as relatively impoverished, at least in comparison with that of humans.

8 We see these motivations come apart for the specific euthanasia cases we present, so they must be distinct. However, we have not shown nor do we claim that compassion is always associated with resistance to euthanasia. In other

unpublished studies using different scenarios we have replicated the resistance finding a few times, and in one case found the reverse association. We are not yet clear on the factors that influence when compassion causes resistance versus approval of euthanasia. We believe it relates to whether it is euthanasia or suffering that is perceived as more dehumanizing. We hypothesize that, if the suffering is perceived to be destroying or ruining the individual's soul, then compassion will drive approval of euthanasia.

9 It is not our claim that moral engagement leads to resistance to harming in all cases. The majority of US citizens are prolife and the majority supports the death penalty. We suspect these attitudes are both driven by moral engagement (moral concern and moral outrage, respectively). Again, we find that religious language appears to best capture these motivations—we seek to protect an innocent soul, but we approve of harm to the individual who has shown that she or he has no soul.

10 Certainly some reviewers of earlier versions of this manuscript were critical of this "gloss" on our findings. Yet they were also at a loss to offer any adequate alternative description.

11 The utilitarian may of course attempt to resist this, dismissing such prosocial personality traits as undesirable because they don't lead to utilitarian thinking, and countering that our argument is circular because we are relying on implicit normative judgments. We wish them well with this argumentative strategy. Ethics is ultimately a practical business, and hence the "utility" of a given ethical framework depends on facts about human psychology—including the fact that psychopathic tendencies decrease the happiness of others and in extreme cases cause people to have little compunction about harming others. In principle it is an empirical matter to determine which ethical system it is best to adopt in order to maximize utility (however that is measured). This will in turn influence which personality traits attract social approval, are actively cultivated, and determine the social power of individuals. Extrapolating from the current evidence, it seems plausible that the act utilitarian will be defeated on her own terms because her ethical framework will be shown to have low utility. Rule utilitarianism is a different matter. We hazard the guess that a utility maximizing version of rule utilitarianism will determine that individuals do best to adopt (something similar to) deontological principles to guide their personal moral decisions. We find this an intellectually appealing solution to the unavoidable problem of balancing the tensions that arise between analytic and empathetic reason.

References

Amodio, D. M. and Frith, C. D. (2006), "Meeting of minds: The medial frontal cortex and social cognition." *Nature Reviews Neuroscience*, 7(4), 268–77.

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., and Toney, L. (2006), "Using self-report assessment methods to explore facets of mindfulness." *Assessment*, 13, 27–45.

Bagozzi, R. P., Verbeke, W. J. M. I., Dietvorst, R. C., Belschak, F. D., van den Berg, W. E., and Rietdijk, W. J. R. (2013), "Theory of mind and empathic explanations of Machiavellianism: A neuroscience perspective." *Journal of Management*, doi: 10.1177/0149206312471393

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001), "The 'Reading the Mind in the Eyes' Test Revised Version: A study with normal adults, and adults with Asperger Syndrome or High-functioning Autism." *Journal of Child Psychology and Psychiatry*, 42(2), 241–51.

Bartels, D. and Pizarro, D. A. (2011), "The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas." *Cognition*, 121, 154–61.

Batson, D. (2009), "These things called empathy: Eight related but distinct phenomena," in J. Decety and W. Ickes (eds), *The Social Neuroscience of Empathy*. Cambridge, MA: MIT Press, pp. 3–15.

Berker, S. (2009), "The normative insignificance of neuroscience." *Philosophy and Public Affairs*, 37(4), 293–329.

Bloom, P. and Weisberg, D. S. (2007), "Childhood origins of adult resistance to science." *Science*, 316, 996–97.

Boyatzis, R. E. and Goleman, D. (2007), *Emotional and Social Competency Inventory*. Boston: Hay Group.

Buckner, R. and Carroll, D. (2007), "Self-projection and the brain." *Trends in Cognitive Sciences*, 11, 49–57.

Buss, A. H. and Perry, M. P. (1992), "The aggression questionnaire." *Journal of Personality and Social Psychology*, 63, 452–9.

Choe, S. Y. and Min, K-H. (2011), "Who makes utilitarian judgments? The influences of emotions on utilitarian judgments." *Judgment and Decision Making,* 6, 580–92.

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., and McDermott, K. B. (2009), "The contributions of prefrontal cortex and executive control to deception: Evidence from activation likelihood estimate meta-analyses." *Cerebral Cortex*, 19, 1557–66.

Cikara, M., Eberhardt, J. L., and Fiske, S. T. (2011), "From agents to objects: Sexist attitudes and neural responses to sexualized targets." *Journal of Cognitive Neuroscience*, 23(3), 540–51.

Conway, P. and Gawronski, B. (2013), "Deontological and utilitarian inclinations in moral decision making: A process dissociation approach." *Journal of Personality and Social Psychology*, 104, 216–35.

Côté, S., Piff, P. K., and Willer, R. (2013), "For whom do the ends justify the means? Social class and utilitarian moral judgment." *Journal of Personality and Social Psychology*, 104(3), 490–503.

Cushman, F. and Greene, J. D. (2011), "Finding faults: How moral dilemmas illuminate cognitive structure." *Social Neuroscience*, 7, 269–79.

Cushman, F. and Young, L. (2011), "Patterns of moral judgment derive from non-moral psychological representations." *Cognitive Science*, 35(6), 1052–75.

Davis, M. H. (1980), "A multidimensional approach to individual differences in empathy." *JSAS Catalog of Selected Documents in Psychology*, 10, 85.

De Waal, F. (2009), *The Age of Empathy: Nature's Lessons for a Kinder Society*. New York: Random House.

Eisenberger, N. I. and Cole, S. W. (2012), "Social neuroscience and health: Neurophysiological mechanisms linking social ties with physical health." *Nature Neuroscience*, 15(5), 669–74.

Evans, J. S. B. and Frankish, K. E. (2009), *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press.

Formosa, P. (2011), "A life without affects and passions: Kant on the duty of apathy." *Parrhesia: A Journal of Critical Philosophy*, 13, 96–111.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., and Raichle, M. E. (2005), "The human brain is intrinsically organized into dynamic, anticorrelated functional networks." *Proceedings of the National Academy of Science USA*, 102(27), 9673–8.

Fransson, P. and Marrelec, G. (2008), "The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis." *Neuroimage*, 42(3), 1178–84.

Frederick, S. (2005), "Cognitive reflection and decision making." *Journal of Economic Perspectives*, 19, 25–42.

Gaita, R. (1999), *A Common Humanity: Thinking about Love and Truth and Justice*. Melbourne: Text Publishing.

Gleichgerrcht, E. and Young, L. (2013), "Low levels of empathic concern predict utilitarian moral judgment." *PLoS ONE*, 8(4), e60418.

Glenn, A. L. (2011), "The other allele: Exploring the long allele of the serotonin transporter gene as a potential risk factor for psychopathy: A review of the parallels in findings." *Neuroscience and Biobehavioral Reviews*, 35, 612–20.

Gobbini, M. I. and Haxby, J. V. (2007), "Neural systems for recognition of familiar faces." *Neuropsychologia*, 45(1), 32–41.

Gray, K., Young, L., and Waytz, A. (2012), "Mind perception is the essence of morality." *Psychological Inquiry*, 23(2), 101–24.

Greene, J. D. (2007), "The secret joke of Kant's soul," in W. Sinnott-Armstrong (ed.), *Moral Psychology, Vol. 3: The Neuroscience of Morality*. Cambridge, MA: MIT Press, pp. 35–79.

—(2009), "The cognitive neuroscience of moral judgment," in M. S. Gazzaniga (ed.), *The Cognitive Neurosciences IV*. Cambridge, MA: MIT Press, pp. 987–1002.

—(2010), "Does moral action depend on reasoning?" in Big Questions essay series, John Templeton Foundation: http://www.templeton.org/reason/Essays/greene.pdf

—(2012), "Reflection and reasoning in moral judgment." *Cognitive Science*, 36, 163–77.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2009), "Pushing moral buttons: The interaction between personal force and intention in moral judgment." *Cognition*, 111, 364–71.

Greene, J. D., Morelli, S. A., Kowenberg, K., Nystrom, L. E., and Cohen, J. D. (2008), "Cognitive load selectively interferes with utilitarian moral judgment." *Cognition*, 107, 1144–54.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004), "The neural bases of cognitive conflict and control in moral judgment." *Neuron*, 44, 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001), "An fMRI investigation of emotional engagement in moral judgment." *Science*, 293, 2105–8.

Gross, J. J. and John, O. P. (1997), "Revealing feelings: Facets of emotional expressivity in self-reports, peer ratings, and behavior." *Journal of Personality and Social Psychology*, 72, 435–48.

—(2003), "Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being." *Journal of Personality and Social Psychology*, 85, 348–62.

Harris, L. T. and Fiske, S. T. (2006), "Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups." *Psychological Science*, 17(10), 847–53.

Haslam, N. (2006), "Dehumanization: An integrative review." *Personality and Social Psychology Review*, 10, 252–64.

Hill, J., Inder, T., Neil, J., Dierker, D., Harwell, J., and Van Essen, D. (2010), "Similar patterns of cortical expansion during human development and evolution." *Proceedings of the National Academy of Sciences*, 107(29), 13135–40.

Jack, A. I. (forthcoming), "A scientific case for conceptual dualism: The problem of consciousness and the opposing domains hypothesis," in J. Knobe, T. Lombrozo, and S. Nichols (eds), *Oxford Studies in Experimental Philosophy* (Vol. 1). Oxford: Oxford University Press.

Jack, A. I. and Robbins, P. (2012), "The phenomenal stance revisited." *Review of Philosophy and Psychology*, 3(3), 383–403.

Jack, A. I., Dawson, A., Begany, K., Leckie, R. L., Barry, K., Ciccia, A., and Snyder, A. (2013a), "fMRI reveals reciprocal inhibition between social and physical cognitive domains." *NeuroImage*, 66, 385–401.

Jack, A. I., Dawson, A., and Norr, M. E. (2013b), "Seeing human: Distinct and overlapping neural signatures associated with two forms of dehumanization." *NeuroImage*, 79, 313–28.

Jack, A. I., Taylor, S., Boyatzis, R., and Friedman, J. (under review), "A novel dual-process account of religious belief."

Kahneman, D. (2003), "A perspective on judgment and choice." *American Psychologist*, 58, 697–720.

Kant, I. (1785/2005), *Groundwork for the Metaphysics of Morals*. Toronto: Broadview Press.

—(1797/1996), *The Metaphysics of Morals*. Cambridge: Cambridge University Press.

Kinderman, P., Dunbar, R., and Bentall, R. (1998), "Theory-of-mind deficits and causal attributions." *British Journal of Psychology*, 89, 191–204.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., and Damasio, A. (2007), "Damage to the prefrontal cortex increases utilitarian moral judgments." *Nature*, 446, 908–11.

Koenigs, M., Kruepke, M., Zeier, J., and Newman, J. P. (2012), "Utilitarian moral judgment in psychopathy." *Social, Cognitive, and Affective Neuroscience*, 7, 708–14.

Kunzendorf, R. G., Moran, C., and Gray, R. (1996), "Personality traits and reality-testing abilities, controlling for vividness of imagery." *Imagination, Cognition, and Personality*, 15, 113–31.

Lawson, J., Baron-Cohen, S., and Wheelwright, S. (2004), "Empathising and systemising in adults with and without Asperger syndrome." *Journal of Autism and Developmental Disorders*, 34, 301–10.

Leyens, J.-P., Rodriguez-Perez, A., Rodriguez-Torres, R., Gaunt, R., Paladino, M.-P., Vaes, J., and Demoulin, S. (2001), "Psychological essentialism and the attribution of uniquely human emotions to ingroups and outgroups." *European Journal of Social Psychology*, 31, 395–411.

Marsh, A. A., Crowe, S. L., Yu, H. H., Gorodetsky, E. K., Goldman, D., and Blair, R. J. R. (2011), "Serotonin transporter genotype (5-HTTLPR) predicts utilitarian moral judgments." *PLoS ONE*, 6, e25148.

McMahan, J. (2002), *The Ethics of Killing: Problems at the Margins of Life*. New York: Oxford University Press.

Meharabian, A., Young, A. L., and Sato, S. (1988), "Emotional empathy and associated individual differences." *Current Psychology*, 7, 221–40.

Mill, J. S. (1861/1998), *Utilitarianism*. Oxford: Oxford University Press.

Mitchell, J. P. (2009), "Social psychology as a natural kind." *Trends in Cognitive Science*, 13(6), 246–51.

Moll, J. and de Oliveira-Souza, R. (2007), "Moral judgments, emotions and the utilitarian brain." *Trends in Cognitive Science*, 11(8), 319–21.

Nowicki, S. and Duke, M. P. (2001), "Nonverbal Receptivity: The Diagnostic Analysis of Nonverbal Accuracy (DANVA)," in J. Hall and F. Bernieri (eds), *Interpersonal Sensitivity: Theory and Measurement*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 185–204.

Ochsner, K. N., Knierim, K., Ludlow, D. H., Henelin, J., Ramachandran, R., Glover, G., and Mackey, S. C. (2004), "Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other." *Journal of Cognitive Neuroscience*, 16(10), 1746–72.

Paulhus, D. L., Neumann, C. S., and Hare, R. D. (forthcoming), *Manual for the Self-Report Psychopathy Scale*. Toronto: Multi-Health Systems.

Radiolab (2007), "Chimp fights and trolley rides." WNYC radio, 13 August 2007. www.radiolab.org/2007/aug/13/chimp-fights-and-trolley-rides/

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001), "A default mode of brain function." *Proceedings of the National Academy of Sciences*, 98, 676–82.

Ritter, R. S., Preston, J. L., and Hernandez, I. (2013), "Happy tweets: Christians are happier, more socially connected, and less analytical than atheists on Twitter." *Social Psychological and Personality Science*, doi: 10.1177/1948550613492345

Robbins, P. and Jack, A. I. (2006), "The phenomenal stance." *Philosophical Studies*, 127, 59–85.

Roy, M., Shohamy, D., and Wager, T. D. (2012), "Ventromedial prefrontal-subcortical systems and the generation of affective meaning." *Trends in Cognitive Sciences*, 16(3), 147–56.

Russell, B. (1926), *Education and the Good Life*. New York: Boni and Liveright.

Saxe, R. and Kanwisher, N. (2003), "People thinking about thinking people: The role of the temporo-parietal junction in 'theory of mind'." *NeuroImage*, 19(4), 1835–42.

Smith, D. L. (2011), *Less Than Human: Why We Demean, Enslave, and Exterminate Others*. New York: St. Martin's Press.

Spunt, R. P. and Lieberman, M. D. (2013), "The busy social brain: Evidence for automaticity and control in the neural systems supporting social cognition and action understanding." *Psychological Science*, 24(1), 80–6.

Strawson, P. F. (1962), "Freedom and resentment." *Proceedings of the British Academy*, 48, 1–25.

Terbeck, S., Kahane, G., McTavish, S., Savulescu, J., Levy, N., Hewstone, M., and Cowen, P. J. (2013), "Beta adrenergic blockade reduces utilitarian judgment." *Biological Psychology*, 92, 323–8.

Thomson, J. J. (1985), "The trolley problem." *Yale Law Journal*, 94, 1395–415.

Tillem, S. and Jack, A. I. (2013), "ALE based meta-analyses of social, mnemonic, and action related processes support the opposing domains hypothesis." Poster presented at Annual Meeting of the Association for Psychological Science, May 2013, Washington, DC.

Van Essen, D. C. and Dierker, D. L. (2007), "Surface-based and probabilistic atlases of primate cerebral cortex." *Neuron*, 56(2), 209–25.

Van Overwalle, F. (2009), "Social cognition and the brain: A meta-analysis." *Human Brain Mapping*, 30(3), 829–58.

Wiech, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., and Tracey, I. (2013), "Cold or calculating? Reduced activity in the subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment." *Cognition*, 126, 364–72.

Williams, B. (1973), "A critique of utilitarianism," in J. J. Smart and B. Williams (eds), *Utilitarianism: For and Against*. Cambridge: Cambridge University Press, pp. 77–150.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., and Saxe, R. (2010), "Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments." *Proceedings of the National Academy of Sciences*, 107(15), 6753–8.

Zahn, R., de Oliveira-Souza, R., Bramati, I., Garrido, G., and Moll, J. (2009), "Subgenual cingulate activity reflects individual differences in empathic concern." *Neuroscience Letters*, 457(2), 107–10.

Zajonc, R. B. (1998), "Emotions," in D. T. Gilbert and S. T. Fiske (eds), *The Handbook of Social Psychology*. New York: McGraw-Hill, pp. 591–632.

7

How Many of Us Are There?

Hannah Tierney, Chris Howard, Victor Kumar,
Trevor Kvaran, and Shaun Nichols

In trying to chart the contours of our folk conceptions, philosophy often proceeds with an assumption of monism. One attempts to provide a single account of the notion of free will, reference, or the self. The assumption of monism provides an important constraint for theory building. And it is a sensible starting assumption. Monistic views often provide the simplest explanations with the fewest commitments. However, it's possible that for some philosophically interesting notions, people operate with multiple different notions. Experimental philosophy provides new tools for exploring such folk pluralism.

We will argue that in the case of personal identity, monism does not capture folk commitments concerning personal identity. Many of our identity-related practical concerns seem to be grounded in distinct views of what is involved in personal identity. Furthermore, both empirical evidence and philosophical thought experiments indicate that judgments about personal identity are regimented by two (or more) different criteria.[1] Of course, just because people are pluralists about personal identity doesn't mean that this is a sustainable position. In the second half of the chapter, we will consider reasons for thinking that the folk commitment to pluralism should be rejected or overhauled. We will offer a tentative case in favor of a pluralist philosophical view about personal identity.

1 Folk pluralism about personal identity

Williams (1970) provides one of the most famous sets of thought experiments in contemporary philosophy. He first gives a case involving two persons, A and B.

A is about to have all of his psychological characteristics transferred into B's body, and simultaneously, B's psychological characteristics will be transferred to A's body. Before the transfers, A and B are told that one of the resulting persons will be tortured. Now, Williams asks, before the transfer, should A want the torture to be administered to his current body or to B's current body? Suppose A asks for the torture to be administered to the A-body and B asks for the torture to be administered to the B-body. Given the transfer of psychological characteristics, the person in the B-body will remember A's request. After running through various possible requests and outcomes, Williams writes, "all the results suggest that the only rational thing to do, confronted with such an experiment, would be to identify oneself with one's memories, and so forth, and not with one's body" (1970, p. 167). Williams thus suggests that this thought experiment indicates that what matters for personal identity is one's *psychological characteristics*. Call this the "trait view" of personal identity.

Next, Williams has us consider the same case described slightly differently. Someone tells me that I am going to be tortured tomorrow, but before the torture, all of my distinctive psychological traits will be removed, and false memories will be inserted. After all that, the torture will be administered. What reaction should I have to the torture?

> Fear, surely, would . . . be the proper reaction: and not because one did not know what was going to happen, but because in one vital respect one did know what was going to happen—torture, which one can indeed expect to happen to oneself, and to be preceded by certain mental derangements as well. (1970, p. 168)

On Williams's view, in this case, despite the annihilation of one's psychological characteristics, it seems like one will still be present to feel the pain. Williams maintains that our intuitions about this case conform to a *biological* approach to personal identity. Although the distinctive psychological characteristics are eradicated, the biological organism persists.[2]

Williams's two cases seem to support starkly different views of personal identity. Sider notes that one natural reaction to these cases is to think that people operate with two notions of persistence for persons:

> It appears that we are capable of having either of two intuitions about the case. . . . A natural explanation is that ordinary thought contains

> two concepts of persisting persons, each responsible for a separate set of intuitions, neither of which is *our* canonical conception to the exclusion of the other . . . (Sider 2001, p. 198)

In the following section, we will argue that empirical research indicates that people really do have two different ways of thinking about personal identity, one in terms of psychological traits and another that conforms more closely to a biological criterion.[3]

1.1 Empirical case for pluralism about personal identity

1.1.1 Survey studies

In survey studies, people deploy a trait conception of self in some contexts but not others. For instance, Rips and colleagues presented participants with a scenario in which Jim's brain was placed in a robot, and they varied whether the brain retained Jim's memories. Participants tended to judge that it was still *Jim* only when the brain retained Jim's memories (Rips et al. 2006). In another study, participants were asked directly to list what they regarded as essential for continued existence. Participants were told:

> One problem that philosophers wonder about is what makes a person the *same person* from one time to another. For instance, what is required for some person in the future to be the *same person* as you? What do you think is required for that?

Over 70 percent of the participants explicitly mentioned psychological factors like memory or personality traits as necessary for persistence. All participants were then pressed with a specific question about the necessity of memory: "In order for some person in the future to be *you*, that person doesn't need to have any of your memories." Over 80 percent disagreed with the claim that they could be identical to someone in the future who didn't have any of their memories (Nichols and Bruno 2010).

Although some surveys trigger responses that fit with a trait conception of self, other studies point to something different. In an adaptation of the key Williams case, participants were told to imagine that they were about to undergo a surgery that would permanently eliminate their distinctive mental states (including thoughts, memories, and personality traits). Next they were

told that after the surgery the doctors have to administer a series of painful shots. They were then asked to indicate agreement or disagreement with the sentence "*you* will feel the pain." Participants overwhelmingly agreed with this statement, suggesting that they are thinking of the self in a way that persists despite the complete disruption of the trait-self (Nichols and Bruno 2010).

1.1.2 Manipulation studies

The previous studies elicit people's opinions on simple surveys, and these studies indicate that different kinds of questions elicit responses that conform to different ways of thinking about the persistence of self. A different experimental technique starts by manipulating how people think about the self and then explores how this affects judgment and decision-making. Prompted by Parfit's (1971; 1984) work on personal identity, Dan Bartels and colleagues have manipulated how participants think about *psychological connectedness*, which Parfit defines as "the holding of particular direct psychological connections" (1984, p. 206). In several studies, Bartels and Urminsky (2011) manipulated how people think about psychological connectedness with prompts like the following (high-connectedness in brackets):

> Day-to-day life events change appreciably after college graduation, but what changes the most [*least*] between graduation and life after college is the person's core identity. . . . The characteristics that make you the person you are . . . are likely to change radically around the time of graduation [*are established early in life and fixed by the end of adolescence*]. . . . Several studies conducted with young adults before and after college graduation have found large fluctuations in these important characteristics [*have shown that the traits that make up your personal identity remain remarkably stable*]. (Bartels and Urminsky 2011, p. 185)

The first point to make is that these manipulations do affect people's judgments about connectedness of the self. When asked to rate on a scale from 0 to 100 (0 = "I will be completely different in the future"; 100 = "I will be exactly the same in the future"), people in the low-connectedness condition give significantly lower ratings than baseline, and people in the high-connectedness condition give significantly higher ratings than baseline (Bartels and Urminsky 2011; Bartels and Rips 2010; Bartels et al. 2013).

So, the manipulation works to affect people's immediate judgments about the connectedness of the self. Bartels and Urminsky (2011) assigned participants either to the high-connectedness or low-connectedness condition, and then presented them with a temporal discounting task; in this task, participants had to decide whether to forego a smaller reward now in favor of a larger reward later. Bartels and Urminsky found that the manipulation affected how "patient" people were, that is, how willing they were to wait for the larger reward. Those who were given the low-connectedness manipulation were more impatient—more likely to take the smaller reward sooner rather than wait longer for the larger reward, as compared to those who read the high-connectedness manipulation.[4] Subsequent research showed that people given the low-connectedness manipulation are also more charitable with future funds than those given the high-connectedness manipulation (Bartels et al. 2013). Thus, leading people to think that the self changes a lot seems to make them less concerned with the interests of the future self, and relatively more concerned with the interests of others.

We extend this work by exploring whether manipulating people's beliefs about psychological connectedness also affects their judgments about punishment. Half of the participants were in the low-connectedness condition and half were in the high-connectedness condition. Each of these groups was also split into past (1-year) and present (1-week) conditions (yielding a 2×2 design). Following the connectedness manipulation, participants were instructed to: "Imagine that you cheated on an exam a week [*year*] ago." Participants were then asked if their cheating were discovered now, how much punishment would be appropriate (ranging from none at all [1] to the maximum punishment allowed by the school [6]). We expected that if people are led to believe that there is low psychological connectedness, this shouldn't affect their judgments about the allotment of punishment for a very recent offense. Thus, we predicted no difference between high and low-connectedness for participants asked about an offense committed a week ago. But for those asked about an offense committed a year ago, we predicted that those who thought there was low psychological connectedness would regard themselves as deserving less punishment. That is exactly what we found (see Figure 7.1).[5] Intuitively, the idea is that if I think that I'm psychologically very different from my past self, then I will regard my current self as less deserving of punishment.

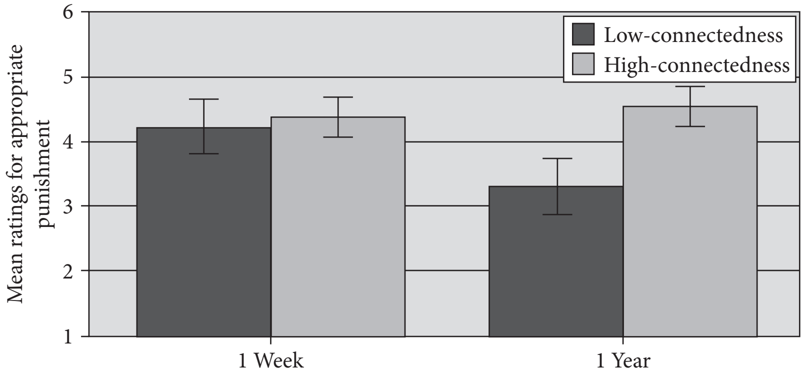


Figure 7.1 Connectedness and deserving punishment (bars represent two standard errors of the mean).

Thus far, we've seen that manipulating beliefs about connectedness affects practical concerns about economic decisions and punishment. Next, we wanted to explore a Williams-style case. Recall that in Williams's key thought experiment, he asks whether you would fear pain that was scheduled to occur after massive psychological changes. Philosophers have tended to follow Williams in saying that one would fear the pain. On this basis, we predicted that anxiety about future pain would *not* be affected by the connectedness manipulation. That is, leading people to think that the self actually changes a lot won't affect their anxiety about future pain. As before, participants were presented with either the high or the low-connectedness manipulation and then given the following instructions:

> Imagine that you have to get a root canal in 1 week [*1 year*]. To what extent would you be anxious about the root canal right now?

While anxiety was greatly affected by whether the pain would occur in a week or a year, connectedness had no significant effect on reported anxiety (see Figure 7.2).[6]

These manipulation studies use a very different technique than the survey studies, but once again indicate that people operate with two different conceptions of self. When people are led to think that their future self won't be very connected with their current self, this affects their economic decisions and punishment judgments. But when it comes to contemplating future pain, manipulating beliefs about connectedness seems to have no effect.

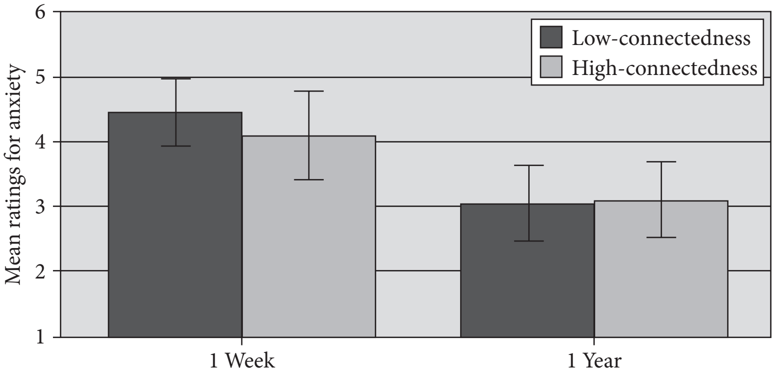


Figure 7.2 Connectedness and anxiety about future pain (bars represent two standard errors of the mean).

Changing the way people think about the self *qua* collection of traits does not impact their feelings about future pain. This, of course, conforms with the results of the simple survey involving a Williams-style case (Nichols and Bruno 2010).

1.1.3 Senses of self

Survey and manipulation studies both indicate that people's judgments conform to different criteria for personal identity in different contexts. Recent work on memory provides a partial explanation for this. There are apparently two very different senses of self implicated in long-term memory. One of these senses of self is stored in *semantic memory* and presents the self as a collection of psychological traits. A series of studies with brain-damaged patients shows that this trait sense of self is preserved across massive psychological damage, including profound retrograde and anterograde amnesia (e.g., Klein et al. 2004; Tulving 1993). Even patients who can't remember a single episode from their past still have knowledge of their personality characteristics. Given that they have knowledge of their traits despite a complete absence of episodic memory, these cases suggest that semantic memory holds the trait conception of the self.

While semantic memory stores a representation of self as a collection of traits, *episodic memory* delivers a sense of self that is not tied to a trait set (Nichols 2014). When I recall an event from my childhood, the representation of the self—that this event happened to *me* at age 12—is

not disrupted by the fact that the child's traits differ radically from my current traits. This holds even for people with significant brain damage. Despite profound neurological dysfunction, when brain-damaged patients recall experiences from their childhood, they give every impression of identifying with that distant and qualitatively different individual (see, e.g., Hilts 1995; Skloot 2004). Episodic memory is constructed such that it can likely produce a sense of identity with some person in the past even if there is *no* psychological continuity with the past person. That's just how episodic memory is built.[7]

The prevailing view in contemporary memory theory is that imagining future experience ("episodic foresight") recruits some of the same mechanisms as recollecting past experience (e.g., Hayne et al. 2011). It's plausible that one commonality between episodic recollection and episodic prospection is in how the self is presented. When I remember an experience from childhood, the trait changes that I've undergone do not interfere with remembering the experience from the first person perspective; similarly, when I imagine sitting in my rocking chair as an octogenarian, the fact that I will have very different traits does not interfere with imagining rocking back and forth. Just as episodic retrospection generates a sense of identity across trait changes, so too in imagining the future, episodic prospection seems to allow us to project ourselves into new experiences without any attention to trait differences. This point applies to the Williams-style cases. It's easy for me to imagine feeling pain in the future without giving a moment's consideration to the stability of my psychological traits. This explains why in survey studies, people tend to say that even after a complete psychological transformation, *they* would feel the pain. This also explains why getting people to think that the self changes dramatically has no effect on their anxiety about future root canals. When imagining future pain, the trait-conception of self takes a backseat.

1.2 Pluralism and practical concerns

In the previous section, we argued that both survey studies and manipulation studies support a pluralist approach to characterizing folk views regarding personal identity. Moreover, work in memory research provides independent reason to think that the self gets represented in two very different ways. One of

these representations presents the self as a set of psychological traits. The other representation is decidedly not defined in terms of psychological traits and fits with a biological or "animalist" conception of self. These two different representations issue different identity criteria.

In a series of important articles, David Shoemaker argues that no single account of personal identity is adequate for the plurality of our practical concerns (Shoemaker 2007, 2011, forthcoming). Thus, Shoemaker is a natural ally for our pluralist agenda. Despite this "pluralism of the practical" (Shoemaker forthcoming, p. 23), Shoemaker does not embrace a pluralistic approach to the criteria for personal identity. Rather, he argues that the concept of *ownership* can ground all of our identity-related practical concerns, without an appeal to personal identity at all. In the following section, we first present Shoemaker's argument for the "pluralism of the practical" and then argue, *contra* Shoemaker, that pluralistic judgments cannot be fully explained with reference to the ownership relation he posits.

1.2.1 Shoemaker on the "Pluralism of the Practical"

Shoemaker argues that our practical concerns, though commonly thought to track the psychological continuity of persons, don't track a single notion of personal identity (Shoemaker forthcoming, p. 23).[8] Shoemaker analyzes several identity-related practical concerns and argues that these concerns cannot be captured by a singular, monistic view of personal identity.

According to Shoemaker, social treatment often tracks biological, rather than psychological, continuity:

> Consider someone who, due to some traumatic brain injury, undergoes radical psychological discontinuity. She will still be treated as the owner of the pre-transformation-person's car and other property, and she will also be treated as the spouse of the pre-transformation-person's spouse, the daughter of her parent, and so forth. (p. 24)

To be sure, we don't think it would be acceptable for the state to seize the property of people suffering from dementia, even for the demented with no heirs. And we don't disapprove of the love expectant mothers feel for their fetuses or children feel for their elderly and vegetative parents. In these cases, social treatment depends on biological, not psychological, continuity (see also

Schechtman 2010, pp. 275–6). Shoemaker also notes that compensation tracks biological continuity in certain situations:

> Suppose Johann suddenly enters a fugue state. Call the radically psychologically discontinuous 'fuguer' Sebastian. Suppose that I had broken Johann's wrist prior to the fugue state but that I now have the medical equipment and expertise to completely heal it and, indeed, make it stronger than before (i.e. to 'rejuvenate' it). When I rejuvenate the wrist I broke, it is Sebastian's. Does what I have done count as compensation? It certainly seems so, despite the psychological discontinuity between Sebastian and Johann. This is because the kind of burden I attempted to rectify was to Johann's animal self . . . (Shoemaker forthcoming, p. 30)

Despite the lack of psychological continuity between Johann and Sebastian, an act of compensation remains possible. In this context, Shoemaker argues that compensation tracks biological continuity as opposed to psychological continuity.

Of course, both social treatment and compensation can also track psychological continuity. If a child's entire psychological profile was transferred from her body to a new body, while the old body was destroyed, her mother would presumably treat this "new" individual as her own child, despite the lack of biological continuity. And, if an individual had caused some psychic trauma to the child before the transfer, but rectified the trauma after the transfer occurred, the child would still surely be compensated, again, despite the lack of biological continuity. In this way, it looks as though these identity-related practical concerns fail to track a singular, monistic criterion of identity, but rather follow two distinct criteria in different contexts.

1.2.2 Identity versus ownership

Shoemaker's cases seem to provide yet further reason to think that judgments about personal identity are pluralistic. However, Shoemaker does not think that this plurality of the practical supports a pluralist theory of personal identity. Indeed, he explicitly rejects pluralism about personal identity (Shoemaker forthcoming, p. 31). Rather, he argues that the concept of *ownership* grounds the plurality of practical concerns, without implicating the concept of identity at all. While it is beyond the scope of this chapter to analyze Shoemaker's ownership relation in full, we are not convinced that this relation can capture

the relevant set of practical judgments without appealing to identity. To illustrate this, we will analyze two of the practical concerns Shoemaker argues are best grounded in terms of the ownership relation, and raise some worries for each.

According to Shoemaker, ascriptions of responsibility can be explained in terms of ownership. We hold agents responsible for their actions because they are *theirs*, not because these agents are identical to the agents who performed them. While it may seem as though ownership and personal identity are intimately connected, Shoemaker notes that "ownership is a relation between agent and action, whereas identity is a relation some individual bears to itself" (Shoemaker forthcoming, p. 25). Though we do not wish to argue that these relations are identical, there is good reason to think that judgments of identity are critical to judgments of ownership. For example, part of what makes a previously performed action *mine* is that I am the same person as the agent who performed the action. Indeed, a common way to deny ownership of (and responsibility for) a given action is to argue that you are not the same person as the agent who performed it.

To prise apart ownership from identity, Shoemaker argues that an agent can own an action without being identical to the agent who performed it. Shoemaker presents two cases in which ownership explains attributions of responsibility yet identity does not obtain.

Shoemaker first points to joint acts, in which two or more agents both own and are responsible for an action, though they are not identical to one another. For example, though Jones and Smith both perform a bank robbery, and are equally responsible for this joint action, neither of them is identical to the "joint agent" that produced the action. But it doesn't follow that identity is irrelevant to these judgments. First, personal identity theorists can admit that multiple agents are capable of being responsible for joint actions without arguing that these agents must be identical to the joint agent. It is enough that an individual is identical to *one* of the people who took part in performing the joint action to establish a judgment of responsibility. More importantly, the judgment that Jones and Smith *own* the action seems to depend on a presupposition of identity. After all, if Jones were not identical to one of the people who helped perform the bank robbery, then we would have good reason to think that Jones does not own the action, and is thus not responsible for it.

Shoemaker then produces cases in which an agent is responsible for an act despite not having performed the act in any way. If a parent tells her child to break a valuable vase, then the parent would be responsible for the child's action even though he is not identical to the child. It seems odd to argue that one could *own* a child's action, even if one is in fact responsible. But in any case, an identity theorist need not argue that an individual is responsible only for those actions he or she performed. Indeed, even if one grants that the ownership relation obtains, it's plausible that it does so in part because the identity relation also obtains. That is, it's because the person who told the child to break the vase was identical to you that we hold you responsible for the action.

Though Shoemaker argues that pluralist intuitions about practical concerns are grounded in terms of ownership, and not in terms of identity, it is far from clear that we can neglect identity in explicating practical concerns. For judgments of ownership seem to depend crucially on judgments about personal identity. Furthermore, the empirical work presented in Section 1.1 provides additional reason to think that assessments of identity really do play a critical role in the formation of judgments regarding our practical interests. The survey tasks explicitly ask about identity (Rips et al. 2006; Nichols and Bruno 2010). The manipulation tasks explicitly manipulate beliefs about identity, and those manipulations affect practical decisions (see also Bartels and Urminsky 2011; Bartels et al. 2013). And the fact that memory systems generate two different senses of the self provides an independent explanation for why people might operate with multiple criteria for identity—the different senses of self ground different identity criteria.

Shoemaker has a further reason to resist pluralism about identity—he maintains that pluralism about personal *identity* is incoherent: "we cannot be pluralists about numerical identity; we can only be pluralists about ownership" (Shoemaker forthcoming, p. 31). One worry is that if there are plural criteria for personal identity, there can be cases that "require that I am both identical with, and not identical with, some past or future individual" (Shoemaker forthcoming, p. 31). More generally, the idea that there are plural criteria for personal identity threatens to generate a problem of too many thinkers. As I deliberate about metaphysics, is the thinker the animal organism or the collection of psychological traits? Are *both* the organism and the psychology engaged in deliberation?

Even if there is some deep incoherence in pluralism about personal identity, it's important to recognize the appropriate target of this concern. It might well be incoherent to *sustain* a pluralist account of identity. But that wouldn't show that people don't *have* a pluralistic conception. As we've seen, there is very good reason to think that people do operate as pluralists about personal identity. Of course, if it really is deeply incoherent to be a pluralist about personal identity, that might give us good reason to reject commonsense pluralism. But the goal of this first part of our chapter was merely to establish that people are committed to plural criteria of personal identity. In the next part of the chapter we examine whether we should abandon those commitments.

2 Philosophical pluralism about personal identity

The fact that individuals' judgments about practical concerns track varying conceptions of personal identity does not entail that we, as philosophers, should endorse pluralism about personal identity. It could be the case that people are simply mistaken when they make judgments about practical concerns that track biological views of personal identity, for example. Perhaps children are unwarranted in feeling obligated to pay for medical care for their senile parents. Alternatively (or even additionally), it could be the case that people are mistaken when they make judgments about practical concerns that track psychological views of personal identity. Perhaps people are unwarranted in attending to psychological continuity or connectedness in allotting punishment. In this part of the chapter, we will explore—in a preliminary and partial way—the philosophical sustainability of pluralism about personal identity.

2.1 The price of monism

Much of philosophy gets its interest precisely because it flows from commonsense conceptions. Accordingly, on many views of metaphysics, folk conceptions define domains (e.g., Lewis 1972; Jackson 1998; Strawson 1959). Jackson defends this view by noting that until we have defined the subject matter, we cannot make progress on metaphysical questions. Jackson uses the analogy of a bounty hunter to make the point. A bounty hunter will "not get

very far if they fail to attend to the representational properties of the handbill" (Jackson 1998, p. 30). "Likewise," Jackson writes,

> Metaphysicians will not get very far with questions like: Are there *K*s? Are *K*s nothing over and above *J*s? and, Is the *K* way the world is fully determined by the *J* way the world is? in the absence of some conception of what counts as a *K* and what counts as a *J* (1998, pp. 30–1)

How are we to determine the subject matter for philosophical issues like the self and free action, then? Jackson maintains that for many philosophical questions, the most plausible answer is one that appeals to ordinary conceptions:

> What . . . are the interesting philosophical questions that we are seeking to address when we debate the existence of free action and its compatibility with determinism . . .? What we are seeking to address is whether free action *according to our ordinary conception* or something close to our ordinary conceptions, exists and is compatible with determinism. (1998, p. 31)

If we take this approach to thinking about the subject matter of the metaphysics of personal identity, it provides a presumption in favor of sustaining pluralism about the *subject matter* of personal identity. That is, the fact that we *have* two senses of self and that they regiment judgments and decisions differently in different contexts provides prima facie reason to think that when we are doing the philosophy of personal identity, we need to adopt a pluralist approach to the subject matter. In some contexts, the ordinary conception of personal identity is trait-based; in other contexts, the ordinary conception of personal identity is biological. Forcing monism on the subject matter would be akin to a bounty hunter searching for one person from the morphed images of two very different people.

It's not just that pluralism fits with what people *say*. As we've seen, pluralism also captures people's *practices*. Our commitments about possessions, inheritance, and familial obligations all conform to the biological conception. To extirpate that conception would be wildly disruptive to our practical concerns. Indeed, it's not clear that it's within the realm of reasonable possibility that we *could* eradicate those kinds of biologically-based practical concerns.

At the same time, people attach enormous importance to their memories and values for personal survival. Furthermore, as we've seen, judgments about punishment are sensitive to psychological continuity and connectedness.

This comports well with the fact that some of the pragmatic aims of punishment, for example, moral education, fail to make sense if our only standard for punishment was a biological criterion. We often punish those who perform illicit actions in order to alter these individuals' psychology such that they recognize the wrongness of their ways, or at least form an aversion to committing future wrongs. Furthermore, as philosophers since Locke have emphasized, what makes an agent responsible for an action seems not simply to be the fact that she is biologically continuous with the agent who performed the action. Rather, we take an agent's psychological characteristics to bear directly on whether she is responsible for an action, and whether punishment is appropriate. Thus, our intuitions about the propriety of punishment and some of the most important justifications of punishment cannot be captured in solely biological terms.

2.2 The coherence of pluralism

In debates over personal identity, animalists bring to bear reasons to reject psychological criteria, and their opponents marshal reasons to reject animalism. Obviously we cannot attempt to address the wide range of arguments brought against one theory or the other. Instead, we want to focus on whether there are good reasons to object to the *pluralistic* aspect of commonsense thought about personal identity. One natural concern is that to embrace a pluralistic approach to personal identity is to embrace blatant inconsistencies. For instance, endorsing the pluralistic approach requires me to say that I both am and am not the same person as a baby born some time ago.

To explore the viability of pluralism, we want to consider how it fares against these kinds of concerns. In the recent literature, pluralism has not been the target of much direct discussion, but the view might seem to be susceptible to one of the most prominent problems raised by animalists: the too-many-thinkers problem. There are many different ways to present the problem. Eric Olson, in his book *What are We?*, sets it up in the following way:

> Suppose human animals think in just the way that we do: every thought of yours is a thought on the part of a certain animal. How could that thinking animal be anything other than you? Only if you are one of at least two beings that think your thoughts? (Or maybe you and the animal think numerically

> different but otherwise identical thoughts. Then you are one of at least two beings thinking exactly similar thoughts.) If you think, and your animal body thinks, and it is not you, then there are two thinkers sitting there and reading this book. (Olson 2007, p. 35)

According to Olson, this is counterintuitive for several reasons. First, it would be very odd if one body housed several different thinkers who thought exactly the same thoughts, yet only one of these thinkers is *you*. However, this is only odd on a monist's view. A pluralist can happily grant that there are two thinkers: under some circumstances you are the animal; under other circumstances, you are the constellation of psychological traits. Of course, one could still argue that it would be odd for two individuals to be housed in one body. The counterintuitiveness of this is acknowledged by four-dimensionalists themselves (e.g., Lewis 1976, p. 26). But the pluralist need not be committed to this kind of multiple-occupancy view. For it's plausible that the animal and the psychological traits do not occupy exactly the same spatio-temporal region.

An epistemic worry also arises from the too-many-thinkers problem: if there are two thinkers housed in your body, how do you know which thinker is you?

> Worse, you *ought* to wonder which of the two thinkers is you. You may believe that you are the non-animal (because you accept the Psychological Approach, perhaps). But the animal has the same grounds for believing that *it* is a non-animal as you have for supposing that you are. Yet it is mistaken. For all you know, you might be the one making this mistake. If you *were* the animal and not the person, you would never be any the wiser. (Olson 2009, p. 82)

Again, such a challenge is only a problem for the monist. While proponents of the psychological approach must address the possibility that there is no way, in principle, to distinguish the biological thinker from the psychological thinker, pluralists are not affected by such a worry. On a pluralist view of personal identity, there is no dispute to settle, for both the biological thinker and the psychological thinker are correct—they *are* both you.

While the too-many-thinkers problem may indeed be a problem, it seems only to be a problem for monists about personal identity. Indeed, even the description of this objection strikes us as an attack on monism, not pluralism—you can only have too many thinkers if you think there should only be one.

However, there is another worry closely related to the too-many-thinkers problem that may challenge the metaphysical coherence of pluralism. Shoemaker writes:

> . . . we cannot be pluralists about numerical identity. . . . If the numerical identity we are talking about is identity of individuals like us, then the proposal just given could require that I am both identical with, and not identical with, some past or future individual. (Shoemaker forthcoming, p. 31)

At first glance, Shoemaker's worry seems deeply troubling: how can it be that an agent is both identical and not identical to the same individual? The pluralist view has a natural response to Shoemaker's concern, however. Consider a case in which an individual is biologically continuous or identical with a past individual, but is not psychologically continuous or identical with that same individual. In one sense, it can be said that the individual both is and is not identical with this past individual, but this phraseology is misleading. It is more apt to say that the individual is biologically identical or continuous with a past individual, but not psychologically identical or continuous. Being biologically continuous and psychologically continuous are two very different states of existence—they require different persistence conditions. If we are explicit about which criterion we are using for our persistence claims, Shoemaker's concern is greatly diminished.

None of the foregoing provides a defense of either the psychological or biological approach to identity. Indeed, for all we have said, it might be the case that ultimately, the right view about personal identity is a nihilistic one on which the self is simply an illusion. There might be deep problems with our ordinary conceptions of the self. But the mere fact that we have multiple conceptions does not look to present us with obvious incoherency.

3 Conclusion: The pull of pluralism

It is a common strategy in experimental philosophy to explain classic philosophical debates—compatibilism versus hard determinism, for example—by isolating different psychological systems that undergird each competing view (see Nichols and Knobe 2007). Often, however, this method only succeeds in explaining the debate, not resolving it, for the two competing views, though

both grounded in psychological systems, remain incompatible. Interestingly, we may be able to do more when it comes to the debate between biological and psychological approaches to personal identity. When people make judgments about persistence and engage in practical decision-making, different contexts seem to trigger different criteria of personal identity. Both empirical evidence and philosophical thought experiments indicate that judgments about personal identity are regimented by two different criteria, one in terms of psychological traits and one that largely conforms to biological criteria. Only a pluralist conception of the self can capture all of our intuitions about persistence—be it those generated by Williams's famous thought experiment or our everyday attributions of moral responsibility. Furthermore, we've argued that some of the most natural objections against pluralism about personal identity do not threaten the kind of pluralism that seems to be implicit in folk judgments and practices. It is possible to synthesize the biological and psychological approach to personal identity into a coherent pluralist view of the self. The monistic presumption that guides philosophical inquiry might be flatly misplaced when it comes to the self. While the assumption of monism is often a sensible theoretical constraint, a monistic view cannot adequately capture folk conceptions of personal identity. Perhaps pluralism has much more pull than originally thought.

Acknowledgments

We'd like to thank Mike Bruno, Wesley Buckwalter, Joshua Knobe, Carolina Sartorio, and Justin Sytsma for comments and discussion on an earlier draft of this chapter. We'd also like to thank David Shoemaker for extensive discussion about the chapter.

Notes

1 In this chapter, our point is to defend pluralism about personal identity. To that end, we focus on two different criteria: psychological and biological. But it could very well be the case that a dual-criteria pluralism isn't plural enough. There might be additional criteria for personal identity that guide judgment in some cases.

2 In keeping with contemporary discussions, we will consider the biological continuity theory as the alternative to the psychological trait theory (Olson 1997; Sider 2001; Shoemaker forthcoming). It's not obvious that the lay intuitions that run against psychological trait theories can be neatly captured by the biological theory. After all, most people believe in survival after death, and that obviously can't be accommodated by the biological theory. It's not clear that beliefs about postmortem survival neatly conform to the trait theory either, though. In some contexts, intuitions about postmortem survival might conform best to the soul theory of personal identity (e.g., Reid 1785). We will have to set those complex issues aside for present purposes.

3 While we talk about conditions of personal identity in this chapter, almost everything we say can be reframed in terms of conditions for "survival" (see e.g., Parfit 1971, p. 8; Rovane 1990, p. 356). Parfit (1971; 1984) argues that fission destroys identity, but does not destroy what matters for survival, and goes on to suggest that what is important in survival is *not* identity but rather *connectedness*. Parfit defines psychological connectedness, as "the holding of particular direct psychological connections" (1984, p. 206; 1971). And this notion is then used to define *psychological continuity*: "the holding of overlapping chains of strong connectedness." As Parfit notes, connectedness is not an identity relation since it isn't transitive (1984, p. 206ff.) Two individuals are psychologically connected if they share direct psychological relations, and this relation can obviously come in degrees (see also Shoemaker 1996, p. 320). In contrast, psychological continuity is an all-or-nothing relation that *is* transitive; it holds between two individuals if chains of strong psychological connectedness obtain between them (Parfit 1984, p. 207). According to Parfit, "Of these two general relations, connectedness is more important both in theory and in practice" (1984, p. 206). Indeed, some of the studies we discuss below focus on the relationship between beliefs about connectedness and judgments about identity-related practical concerns. However, when we discuss the psychological factor in judgments about identity, we intend to be neutral about whether the relevant factor is connectedness or continuity.

4 Bartels and Urminsky also got this effect when they simply measured beliefs about connectedness (Study 5). They had participants indicate how connected they believed they would be with their future self. Three weeks later, they engaged in a temporal discounting task (without being told that this was related to the activity three weeks prior). Bartels and Urminsky found that people who regarded themselves as less connected in part 1 were more impatient on the task in part 2.

5 Consistent with expectations, the analysis revealed the main effects of timing (week vs. year) and connectedness ($Fs(1,181) = 14.83$ and 4.01, $ps < 0.05$) and most importantly, the predicted interaction ($F(1,181) = 9.17$, $p < 0.01$). N = 182, 57 female. All participants were recruited from Amazon Mechanical Turk.

6 There was a main effect of timing [week vs. year] ($F(1, 131) = 17.1$, $p < 0.001$), but there was no effect of connectedness ($F(1,131) = 0.134$, $p = 0.714$) and no interaction ($F(1,131) = 0.266$, $p = 0.607$). N = 132, 61 female. All participants were recruited from Amazon Mechanical Turk.

7 While the conception of self delivered by episodic memory often parallels the biological conception, it is not a perfect fit. For instance, some people report episodic memories of past lives, that is, memories of events that they think they experienced as *other* biological organisms (see, e.g., Spanos et al. 1991). As mentioned above (Footnote 2), we will set aside these complications for the purpose of this chapter.

8 In this section, we write primarily in terms of psychological continuity because Shoemaker sees his target as those who defend continuity, as opposed to connectedness, criteria. But as we noted above (Footnote 3), we wish to remain neutral between these psychological accounts.

References

Bartels, D., Kvaran, T., and Nichols, S. (2013), "Selfless giving." *Cognition,* 129(2), 392–403.

Bartels, D. and Rips, L. (2010), "Psychological connectedness and intertemporal choice." *Journal of Experimental Psychology: General*, 139, 49–69.

Bartels, D. and Urminsky, O. (2011), "On intertemporal selfishness: The perceived instability of identity underlies impatient consumption." *Journal of Consumer Research*, 38, 182–98.

Hayne, H., Gross, J., McNamee, S., Fitzgibbon, O., and Tustin, K. (2011), "Episodic memory and episodic foresight in 3-and 5-year-old children." *Cognitive Development*, 26(4), 343–55.

Hilts, P. (1995), *Memory's Ghost: The Nature of Memory and the Strange Tale of Mr. M*. New York: Simon & Schuster.

Jackson, F. (1998), *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Clarendon Press.

Klein, S., German, T., Cosmides, L., and Rami, G. (2004), "A theory of autobiographical memory: Necessary components and disorders resulting from their loss." *Social Cognition*, 22, 460–90.

Lewis, D. (1972), "Psychophysical and theoretical identifications." *Australasian Journal of Philosophy*, 50(3), 249–58.

—(1976), "Survival and identity," in A. Rorty (ed.), *The Identities of Persons*. Berkeley, CA: University of California Press, pp. 17–40.

Nichols, S. (2014), "The episodic sense of self," in J. D'Arms and D. Jacobson (eds), *Moral Psychology and Human Agency*. Oxford: Oxford University Press.

Nichols, S. and Bruno, M. (2010), "Intuitions about personal identity: An empirical study." *Philosophical Psychology*, 23(3), 293–312.

Nichols, S. and Knobe, J. (2007), "Moral responsibility and determinism: The cognitive science of folk intuitions." *Noûs*, 41, 663–85.

Olson, E. (1997), *The Human Animal: Personal Identity without Psychology*. Oxford: Oxford University Press.

—(2007), *What are We?* Oxford: Oxford University Press.

—(2009), "Personal identity," in S. Schneider (ed.), *Science Fiction and Philosophy: From Time Travel to Superintelligence*. Malden, MA: Blackwell, pp. 67–90.

Parfit, D. (1971), "Personal identity." *The Philosophical Review*, 80(1), 3–27.

—(1984), *Reasons and Persons*. Oxford: Oxford University Press.

Reid, T. (1785), *Essays on the Intellectual Powers of Man*. Edinburgh: J. Bell.

Rips, L., Blok, S., and Newman, G. (2006), "Tracing the identity of objects." *Psychological Review*, 113, 1–30.

Rovane, C. (1990), "Branching self-consciousness." *The Philosophical Review*, 99(3), 355–95.

Schechtman, M. (2010), "Personal identity and the practical." *Theoretical Medicine and Bioethics*, 31, 271–83.

Shoemaker, D. (1996), "Theoretical persons and practical agents." *Philosophy & Public Affairs*, 25(4), 318–32.

—(2007), "Personal identity and practical concerns." *Mind*, 116, 316–57.

—(2011), "Moral responsibility and the self," in S. Gallagher (ed.), *Oxford Handbook of the Self*. Oxford: Oxford University Press, pp. 487–520.

—(forthcoming), "The stony metaphysical heart of animalism," in S. Blatti and P. Snowdon (eds), *Essays on Animalism: Persons, Animals, and Identity*. Oxford: Oxford University Press.

Sider, T. (2001), "Criteria of personal identity and the limits of conceptual analysis." *Philosophical Perspectives*, 15, 189–209.

Skloot, F. (2004), *In the Shadow of Memory*. Lincoln: University of Nebraska Press.

Spanos, N., Menary, E., Gabora, N., DuBreuil, S., and Dewhirst, B. (1991), "Secondary identity enactments during hypnotic past-life regression: A sociocognitive perspective." *Journal of Personality and Social Psychology*, 61, 308–20.

Strawson, P. (1959), *Individuals: An Essay in Descriptive Metaphysics*. New York: Anchor Books.

Tulving, E. (1993), "Self-knowledge of an amnesic individual is represented abstractly," in T. K. Srull and R. S. Wyer (eds), *Advances in Social cognition* (Vol. 5). Hillsdale, NJ: Erlbaum, pp. 1–49.

Williams, B. (1970), "The self and the future." *The Philosophical Review*, 79, 161–80.

8

Concepts: Investigating the Heterogeneity Hypothesis

Edouard Machery

In *Doing without Concepts* (Machery 2009; see also Machery 2005), I argue that the class of concepts divides into several kinds that have little in common, including prototypes, exemplars, and causal theories—a hypothesis called "the heterogeneity hypothesis."[1] On this view, a class (e.g., the class of dogs or the class of cats), a substance (e.g., water or gold), or an event (e.g., going to the dentist or going to a restaurant) is typically represented by several coreferential concepts. For instance, dogs are hypothesized to be represented by three distinct concepts: a prototype of dogs, a set of exemplars of dogs, and a causal theory about dogs. The evidence put forward for this view in *Doing without Concepts* takes two main forms. First, I argue that some experimental findings are best explained if some concepts are prototypes, other experimental findings are best explained if other concepts are sets of exemplars, and yet other findings are best explained if yet other concepts are causal theories. So, the pattern of findings observed in 40 years of experimental research on concepts is best explained by the heterogeneity hypothesis. Second, I allude toward some more direct experimental evidence for the co-occurrence of distinct coreferential concepts. In more recent work (Machery and Seppälä 2010), I have gone beyond theorizing on the basis of existing evidence, and I have started seeking new experimental evidence bearing on the heterogeneity hypothesis. The goal of this chapter is to extend the work done in Machery and Seppälä (2010). I show that a substantial proportion of competent speakers are willing to endorse apparently contradictory sentences such as "Tomatoes are vegetables" and "Tomatoes

are not vegetables." Revising somewhat the polysemy hypothesis put forward by Machery and Seppälä, I conclude that many words can express different coreferential concepts, consistent with the heterogeneity hypothesis put forward in Machery (2009).

Here is how I will proceed. Section 1 sets up the background of the research started in Machery and Seppälä (2010). Section 2 presents a replication of the data presented in Machery and Seppälä (2010). Section 3 examines the role of the hedge "in a sense" in the data presented in Machery and Seppälä (2010) and in Section 2. Section 4 examines whether these data are systematically influenced by demographic variables. Section 5 extends this research with new stimuli.

1 The polysemy hypothesis

1.1 From the heterogeneity hypothesis to the polysemy hypothesis

The polysemy hypothesis states that words, in particular predicates, often have several different meanings because they lexicalize several distinct coreferential concepts. For instance, "dog" (say) may have three different meanings because it lexicalizes a prototype of dogs, a set of exemplars of dogs, and a causal theory about dogs. So, "dog" has several meanings in roughly the way "bank" has several meanings—namely, a place that receives and lends money and the side of a river.

The polysemy hypothesis is not entailed by the heterogeneity hypothesis since it could be the case that only one of the hypothesized coreferential concepts is lexicalized. For instance, even if a prototype of dogs, a set of exemplars of dogs, and a causal theory of dogs really represent the class of dogs, it could still be the case that only one of these concepts—say, the causal theory of dogs—is lexicalized by "dog." On the other hand, if the polysemy hypothesis is true, then this would provide empirical support for the heterogeneity hypothesis.[2]

The polysemy hypothesis could be true for only some words: That is, even if it is typically the case that classes, substances, and events are represented by several coreferential concepts, it could also be that many, or perhaps even

most, words lexicalize only one of the existing coreferential concepts, while only some, or perhaps only few, words lexicalize several of them. On this view, only some words, and perhaps only few, would be polysemous because of the existence of several coreferential concepts. It could also be the case that, even if the heterogeneity hypothesis is true, whether the existence of several coreferential concepts gives rise to polysemy varies across domains: For instance, artifact words could be more likely to lexicalize several of the coreferential concepts of artifacts, and hence to be polysemous, than words referring to biological kinds. Finally, the polysemy hypothesis could also be true for some individuals only. That is, people could vary (perhaps as a function of some demographic variables) in their tendency to lexicalize several of the existing coreferential concepts (still assuming that the heterogeneity hypothesis is correct). In this chapter, I will examine qualitatively whether polysemy varies across domains; I will also examine more formally whether demographic variables systematically influence the disposition to treat words as polysemous.

1.2 How to test the polysemy hypothesis?

Even if some words are really polysemous because of the heterogeneity hypothesis—for example, even if "dog" expresses a prototype of dogs, a set of exemplars of dogs, and a causal theory about dogs—this polysemy should be much less obvious than the polysemy of "bank." "Bank" is obviously polysemous because it refers to two different kinds of things: It is plain that one walks on the side of a river, but not on the place that receives and lends money, and it is plain that one speaks to a teller in the place that receives and lends money, but not on the side of a river. By contrast, according to the polysemy hypothesis, the distinct concepts that would be lexicalized by a given word (e.g., the distinct concepts of dogs lexicalized by "dog") would be coreferential, and so most of what we take to be true when a word expresses a given concept we would also take it to be true when that word expresses a coreferential concept. People may thus not be aware of the polysemy induced by the existence of several coreferential concepts, and, to assess the polysemy hypothesis, we cannot simply rely on our sense, qua competent speakers, of which words are polysemous and of which are not.

Fortunately, we need not rely on this intuitive sense to identify polysemous words since linguists have developed various tests for polysemy that can be appealed to in order to test the polysemy hypothesis (Zwicky and Sadock 1975; Sennett 2011). In particular, when a word is polysemous, it is possible, without contradiction, to assert one thing of its referent and to deny it. For instance, if I were walking along the river, I could assert and deny without contradiction that I am at the bank:

(1) I am at the bank, but I am not at the bank.

Similarly, I can say without contradiction that a novel in electronic format is a book (meaning, roughly, a text) and not a book (meaning a particular kind of physical object):

(2) The novel I just downloaded from Amazon is a book, but it is not a book.

The suggestion, then, is to rely on this contradiction test to identify polysemous words. If the polysemy hypothesis is true, words that apparently are not polysemous can sometimes be used without contradiction to assert one thing of their referent and to deny it. According to the polysemy hypothesis, this phenomenon should occur in at least two circumstances. Some object can be sufficiently close to the prototype of a class (substance, event type)[3] for people to judge that it is a member of this class, while not being a member of this class according to its causal theory. Alternatively, some object can differ so much from the prototype of a class that people judge that it is not a member of this class, while being a member of this class according to its causal theory. In either situation, if the relevant word expresses both the prototype and the causal theory, then people should both deny and assert that this object belongs to this class, and they should do it without thinking that they are contradicting themselves.

An example may help see the point. Consider the following two sentences:

(3) Tomatoes are a vegetable.
(4) Tomatoes are not a vegetable.

If "vegetable" expresses the prototype of vegetables, people should lean toward judging (3) true since tomatoes have many typical properties of vegetables.

On the other hand, if "vegetable" expresses the causal theory of vegetables (according to which, let's suppose, fruits and vegetables are exclusive classes and according to which whatever has seeds is a fruit), people should lean toward judging (4) true. But because (3) and (4) are true when "vegetable" expresses two distinct concepts, people should realize that they are not really contradicting themselves, and, as a result, they should be willing to assert the two sentences at the same time.

1.3 Previous work

Machery and Seppälä (2010) used six pairs of target sentences similar to the pair of sentences (3) and (4) above and three pairs of control sentences for which the hypothesized prototype and the hypothesized causal theory both lead to the same judgment (viz. that the object is an *x*).[4] In some target pairs, the sentence asserting that the object belongs to a class should be judged to be true because of the similarity of this object to the prototype, while in other target pairs, the sentence denying that the object belongs to a class should be judged to be true because of the large difference between this object and the prototype. The target sentences were about different domains (animals, plants, and artifacts).

In their exploratory study, Machery and Seppälä found that people were up to ten times more likely to agree with both sentences of the target pairs than with both sentences of the control pairs. They also found some variation across sentence pairs. For some target pairs, most people were willing to agree with both sentences, while only a minority were willing to agree with both sentences of other pairs. Unfortunately, Machery and Seppälä were unable to provide an explanation of this variation, in part because of the small number of sentences they used.

In addition, Machery and Seppälä controlled for two possible explanations of their findings. They provided evidence that participants did not interpret the sentences of the target pairs metaphorically. They also provided evidence that people did not agree with both sentences of each target pair (e.g., "Tomatoes are vegetables" and "Tomatoes are not vegetables") because they viewed the objects to be categorized (e.g., tomatoes) as marginal cases of the relevant class (e.g., vegetables).

1.4 The present research

The goal of this chapter is to extend Machery and Seppälä's research. I will first replicate the study using the stimuli used by Machery and Seppälä (Section 2). These stimuli involve the hedge "in a sense" (e.g., "In a sense, tomatoes are vegetables" and "In a sense, tomatoes are not vegetables"). Section 3 examines whether the findings presented by Machery and Seppälä depend on this hedge. Section 4 examines whether there is some systematic individual variation in people's disposition to interpret some words in a polysemous manner: In particular, are some personality types more likely to be so disposed? Section 5 extends Machery and Seppälä (2010) by using other stimuli.

Before reporting the results of these four studies, it is worth highlighting some limits of the present research. First, just like Machery and Seppälä (2010), the research reported here is largely qualitative and exploratory. Typically, only descriptive statistics will be reported, and only a small number of stimuli will be used. Second, some theories of concepts that deny the heterogeneity hypothesis can accommodate the findings reported in this chapter and in Machery and Seppälä (2010). In particular, suppose that concepts are prototypes, and that the weights associated with the properties represented by a prototype can vary across contexts so that some property strongly influences the decision to categorize an object as an *x* in one context, but not in other contexts, where other properties have a greater influence (an assumption which is not uncommon in categorization models). This model of concepts can explain why people would assent to both (5) and (6):

(5) Whales are fish.
(6) Whales are not fish.

They would agree with (5) because they would weigh the appearance and ecology of whales heavily when assessing (5), and they would agree with (6) because they would weigh the capacity to bear live young heavily when assessing (6). Another kind of experimental study is needed to distinguish the heterogeneity hypothesis from this theory of concepts.

2 Study 1: Replication of Machery and Seppälä (2010)

2.1 Participants and materials

Participants were 62 adults recruited on Amazon Turk. Two data points were excluded because the surveys were filled from the same IP address, resulting in a sample of 60 participants (mean age = 35.7; range: 20–71; 59% males). Following Machery and Seppälä (2010), participants were presented online with a survey composed of nine pairs of sentences. Each pair was written on a separate webpage. All the pages were similarly organized. Participants read the following instructions on the top of each page:

> On a scale of 1 to 7, '1' indicating that you totally disagree and '7' indicating that you totally agree, to which extent do you agree with the following two claims? Remember that you might agree with both claims, with only one of the two claims, or with none of them.

A first sentence was written below these instructions. For instance, the first sentence of the first pair was "In a sense, tomatoes are vegetables." A 7-point scale, anchored at 1 with "clearly disagree," at 4 with "not sure," and at 7 with "clearly agree," followed this sentence. The negation of the first sentence was written below this scale.[5] For instance, the second sentence of the first pair was "In a sense, tomatoes are not vegetables." The same 7-point scale followed this second sentence. Participants were then given the opportunity to explain their judgment.

All the sentences expressed (positive or negative) classification judgments: They asked whether a class was included in another class. The nine pairs of sentences consisted of six target pairs and three control pairs. The six target pairs were constructed according to one of the two following principles: (1) in four pairs (A, D, E, and H), it was assumed that the members of the extension of the first predicate of the sentence (e.g., "tomatoes") were similar to the hypothesized prototype expressed by the second predicate (e.g., "vegetables"), but did not belong to the extension of this predicate according to the causal theory I hypothesized it also expresses; (2) in the two remaining pairs (B and G), it was assumed that the members of the extension of the first predicate (e.g., "penguins") were dissimilar to the hypothesized prototype expressed by

the second predicate (e.g., "birds"), but did belong to the extension of this predicate according to the causal theory I hypothesized it also expresses. It is useful to illustrate construction principle 1 with pair A. It was assumed that tomatoes are similar to the prototype of vegetables because they have many of the typical properties of vegetables. At the same time, many people believe that tomatoes are essentially fruits, not vegetables. If the polysemy hypothesis is correct, then participants (or a substantial proportion of them) should be willing to assent both to "In a sense, tomatoes are vegetables" and "In a sense, tomatoes are not vegetables." It is also useful to illustrate construction principle 2 with pair B. It was assumed that penguins are dissimilar to the prototype of birds because they have few of the typical properties of birds (they do not fly) and because they have several properties that are atypical of birds (they swim). At the same time, many people believe that penguins are essentially birds. If the polysemy hypothesis is correct, then participants (or a substantial proportion among them) should be willing to assent both to "In a sense, penguins are birds" and "In a sense, penguins are not birds." I call "theoretical sentences" the sentences hypothesized to be judged true on the basis of the theories of the classes at hand (e.g., "In a sense, tomatoes are not vegetables" and "In a sense, penguins are birds") and "prototypical sentences" those sentences hypothesized to be judged true on the basis of the prototypes of the classes at hand (e.g., "In a sense, tomatoes are vegetables" and "In a sense, penguins are not birds").

Every second target pair was followed by a control pair (C, F, and I). It was assumed that participants would judge the first sentence of a control pair to be clearly true and the second sentence to be clearly false, whatever concept (prototype, theory, etc.) they associate with the second predicate of the sentence. For instance, pair C was made of the two following sentences: (C1) "In a sense, lions are animals" and (C2) "In a sense, lions are not animals." Because lions are animals according to our theoretical beliefs about animals and are typical animals, people should judge C1 true and C2 false whatever concept they associate with "lion."

The nine pairs are presented in Table 8.1. In the target pairs, words expressing concepts that belong to different domains (plants, animals, artifacts, human activities) were used. This is meant to reflect the domain-generality of the heterogeneity hypothesis: It is supposed to apply to all conceptual domains.

Table 8.1 Target and control sentences of Study 1 (control pairs in gray shading, theoretical sentences in italics, prototypical sentences in regular fonts)

Pair	First sentence on a given page	Second sentence on a given page
A	In a sense, tomatoes are vegetables	*In a sense, tomatoes are not vegetables*
B	*In a sense, penguins are birds*	In a sense, penguins are not birds
C	In a sense, lions are animals	In a sense, lions are not animals
D	In a sense, whales are fish	*In a sense, whales are not fish*
E	In a sense, a piano is a piece of furniture	*In a sense, a piano is not a piece of furniture*
F	In a sense, a triangle is a geometric figure	In a sense, a triangle is not a geometric figure
G	*In a sense, chess is a sport*	In a sense, chess is not a sport
H	In a sense, zombies are alive	*In a sense, zombies are not alive*
I	In a sense, a hammer is a tool	In a sense, a hammer is not a tool

Finally, the nine pairs of sentences were followed by a biographical questionnaire. In contrast to Machery and Seppälä (2010), this biographical questionnaire included a personality inventory—the Ten Item Personality Measure (Gosling et al. 2003).

2.2 Data analysis and results

Scoring and data analysis followed Machery and Seppälä (2010). The percentage of participants who gave an answer superior or equal to 4 for both sentences was computed. For each target pair, those participants who gave an incorrect answer to the theoretical sentence were eliminated. For instance, in pair A, the answers of those participants who answered negatively (an answer lower than 4) to the sentence "In a sense, tomatoes are not vegetables" (A2) were ignored, when the score for the dependent measure was computed, since the question of interest is the following one: When people know that a tomato does not fulfill the theory of a vegetable, are they still willing to assent to the sentence, "In a sense, tomatoes are vegetables"? To examine this question, one should only examine the answers of those participants who know that a tomato is (in a sense!) not a vegetable. Generally, in pairs A, D, E, and H, only the answers of those participants who have acquired the belief that although *x*'s look like

Table 8.2 Results of Study 1 (control pairs in gray shading)

	A	B	C	D	E	F	G	H	I
Percentage of agreement with both sentences (≥4)	57.5	30.9	10.3	41.5	59.5	15.2	52.9	49.1	10.9

Table 8.3 Averaged results of Study 1

	Average percentage of agreement with both sentences (≥4)
Target pairs	48.6
Control pairs	12.1

y's, they are *z*'s were of interest: What is at stake is whether they would also be willing to agree that, in a sense, *x*'s are *y*'s. Similarly, in pairs B and G, only the answers of those participants who have acquired the belief that although *x*'s do not look like *y*'s, they are *y*'s were of interest: What is at stake is whether they would be also willing to agree that, in a sense, *x*'s are not *y*'s.

The results for each pair are presented in Table 8.2, and the averaged results across the target and control pairs in Table 8.3.

2.3 Discussion

The results of Study 1 replicate the main qualitative feature of the results reported in Machery and Seppälä (2010): Participants are much more likely to agree with the two sentences of the target pairs than with the two sentences of the control pairs (4 times in Study 1 of this chapter). The variation across pairs is also similar to the variation observed by Machery and Seppälä: The penguin pair is the least likely to elicit agreement to both sentences in this study as was the case in two of the three studies reported by Machery and Seppälä, while the whale pair is the second least likely to elicit agreement to both sentences, as was also the case in two of the three studies reported in Machery and Seppälä. Finally, as was the case in Machery and Seppälä (2010) too, this variation is not easy to interpret.

3 Study 2: The role of "In a Sense"

3.1 Aim of Study 2

The aim of Study 2 is to examine whether the results reported in Machery and Seppälä (2010) and in Study 1 depend on the hedge "in a sense." To examine this question, Study 1 was replicated with the following change in the experimental design: While the stimuli in Study 1 have the structure, "In a sense, *x*'s are (are not) *y*'s," the stimuli in Study 2 have the structure, "*x*'s are (are not) *y*'s."

3.2 Participants and materials

Participants were 42 adults recruited on Amazon Turk (mean age = 34; range: 20–70; 41.5% males). No data point was excluded. Except for the difference mentioned in Section 3.1, the design of Study 2 is identical to the design of Study 1.

3.3 Data analysis and results

Scoring and data analysis were identical to those of Study 1. The results for each pair are presented in Table 8.4, and the averaged results across the target and control pairs in Table 8.5.

Table 8.4 Results of Study 2 (control pairs in gray shading)

	A	B	C	D	E	F	G	H	I
Percentage of agreement with both sentences (≥4)	17.3	8.1	0	6.5	20.7	2.4	25.9	20	0

Table 8.5 Averaged results of Study 2

	Average percentage of agreement with both sentences (≥4)
Target pairs	16.4
Control pairs	0.8

3.4 Discussion

In Study 2 as in Study 1, participants are much more likely (20 times) to agree with the two sentences of the target pairs than with the two sentences of the control pairs. Furthermore, pairs B and D are again the least likely to elicit agreement to both sentences. On the other hand, the results of Study 2 differ dramatically from those of Study 1 in that participants turn out to be much less likely—on average three times less likely—to agree with both sentences of the target pairs. Study 2 provides evidence for the importance of the hedge "in a sense" in the current study: Most people are simply unwilling to agree with two apparently contradictory sentences when a hedge is not used.

These results should lead to a revision of the polysemy hypothesis. The importance of hedges in eliciting assent to apparently contradictory sentences suggests that most words preferentially express a particular concept. For instance, "fish" may preferentially express a causal theory of fish. So, words do not typically express several coreferential concepts, and they are not typically polysemous. But they *can* express several coreferential concepts, provided that a hedge nudges speakers to substitute a coreferential concept for the concept a word preferentially expresses. For instance, if "fish" preferentially expresses a causal theory of fish, competent speakers can also express a prototype of fish by this term. Importantly, when this substitution occurs, "fish" is not used metaphorically, as Machery and Seppälä have shown and as the comments left by some participants to Studies 1 and 2 confirm. For instance, some people who agreed that tomatoes are vegetables gave the following kind of comments:

> Tomatoes are vegetables in the culinary sense, usually how they are prepared in recipes.
>
> Everyone thinks of them as vegetables. why would you want to put fruit sauce on your pizza?

Some people who agreed that whales are fish gave the following kind of comments:

> They swim in the ocean.
>
> They do live in the ocean and have flippers and swim.
>
> Well, yes, they live, travel, give birth, etc. in water, like fish.

This revised interpretation of the findings of Studies 1 and 2 provides support for the heterogeneity hypothesis since it assumes that there are several coreferential concepts that are available to be expressed by a given word.

Perhaps one could object that the most conservative interpretation of the findings in Studies 1 and 2 is that the heterogeneity hypothesis is only true for a minority of individuals. According to this objection, most people are unwilling to assent to apparently contradictory sentences even when a hedge like "in a sense" is used because they do not have several coreferential concepts. Instead, I am proposing that the heterogeneity hypothesis is true of most people, but that people vary in their disposition to substitute a coreferential concept for the concept preferentially expressed by a word. While the data are equally consistent with both interpretations, the interpretation I prefer is more plausible since it would be really surprising if a basic feature of cognitive architecture—what kinds of concepts people have—varied across individuals. It would be less surprising if, for whatever reason, people varied in their willingness to replace a preferred interpretation of a word with a related, but distinct, interpretation.

4 Study 3: Individual variation

4.1 Aim of Study 3

The aim of Study 3 is to examine whether there is some systematic individual variation in people's willingness to assent to two apparently contradictory sentences. I will examine two kinds of factors: gender and the five personality traits identified by the OCEAN model—openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Conscientiousness is of the greatest interest among these five traits since it relates to the tendency to be highly organized, committed, and thorough. It is manifested in a neat and orderly lifestyle. One can speculate that people who are highly conscientious are less likely to be willing to assent to both sentences of the target pairs since these two sentences are apparently contradictory.

To examine whether there is any systematic individual variation, I will use the data set of Study 1 and compare the proportion of agreement to both sentences of the target pairs across personality traits and gender.

4.2 Data analysis and results

For each of the five personality traits, the median value was computed for the total sample size (N = 60), and participants were divided into two groups (e.g., participants low vs. high on openness to experience) on the basis of these five median values. For each personality trait and for gender, the proportion of participants who were willing to agree with both sentences of the target pairs was compared across the two groups.

A few methodological features should be noted. Because the sample size is small, the power of the tests reported in Study 3 is low, and negative results should not be interpreted. Second, although 36 chi-square tests are computed in what follows, the significance level was not adjusted for two reasons (it was left at its conventional value of 0.05). First, the study is largely exploratory: I am interested in finding out whether there may be some personality-related variation in people's willingness to assent to two apparently contradictory sentences. Adjusting the significance level would substantially increase the risk of committing a type-II error. Second, I will conclude that personality or gender has a systematic effect on people's willingness to assent to two apparently contradictory statements only if the same personality trait (e.g., conscientiousness) or gender repeatedly influences people's judgments.

Table 8.6 reports the results of Study 3. Only results significant at 0.05 are reported. The first number reports the proportion of people willing to assent to both sentences in the "low" group (e.g., low in openness to experience), the second number in the high group (e.g., high in openness to experience) when the results are significant.

Table 8.6 Results of Study 3 (ns: nonsignificant)

Pair		A	B	D	E	G	H
Demographic variables	Gender	ns	ns	ns	ns	ns	ns
	Openness	ns	ns	52/20	ns	ns	ns
	Conscientiousness	ns	ns	ns	ns	ns	66/33
	Extraversion	ns	ns	ns	ns	ns	ns
	Agreeableness	ns	ns	ns	ns	ns	69/28
	Neuroticism	ns	ns	ns	48/80	ns	62/32

4.3 Discussion

While the effect of some personality traits reached significance for some target pairs, this effect was not systematic. Particularly, conscientiousness failed to influence people's judgments. Thus, Study 3 failed to find evidence for a systematic effect of people's gender or personality on their willingness to embrace two apparently contradictory sentences, and no evidence was found that some factor systematically influences the likelihood of treating words such as "vegetables" or "dogs" as polysemous. If there is variation, it is either nonsystematic, or it depends on other demographic variables, or its effect size is too small to be detected with the power of Study 3.

5 Study 4: New stimuli

5.1 Aim and design of Study 4

The aim of Study 4 is to extend the research presented in this chapter with new stimuli. Six new target pairs were formulated (see Table 8.7). In every other respect, the design is identical to that of Study 1.

Table 8.7 Target and control sentences of Study 4 (control pairs in gray shading, theoretical sentences in italics, prototypical sentences in regular fonts)

Pair	First sentence on a given page	Second sentence on a given page
A	In a sense, bats are birds	*In a sense, bats are not birds*
B	*In a sense, dolphins are mammals*	In a sense, dolphins are not mammals
C	In a sense, lions are animals	In a sense, lions are not animals
D	In a sense, eels are snakes	*In a sense, eels are not snakes*
E	In a sense, margarine is butter	*In a sense, margarine is not butter*
F	In a sense, a triangle is a geometric figure	In a sense, a triangle is not a geometric figure
G	*In a sense, artist is a job*	In a sense, artist is not a job
H	In a sense, Splenda is sugar	*In a sense, Splenda is not sugar*
I	In a sense, a hammer is a tool	In a sense, a hammer is not a tool

5.2 Participants

Participants were 64 adults recruited on Amazon Turk (mean age = 40.8; range: 18–72; 54.8% males). No data point was excluded.

5.3 Data analysis and results

Scoring and data analysis were identical to those of Study 1. The results for each pair are presented in Table 8.8, and the averaged results across the target and control pairs in Table 8.9.

The analysis of personality traits and gender followed the procedure described in Study 3. Results are reported in Table 8.10.

Table 8.8 Results of Study 4 (control pairs in gray shading)

	A	B	C	D	E	F	G	H	I
Percentage of agreement with both sentences (≥4)	16.4	15.3	0	25.5	22.6	1.6	30.0	31.0	0

Table 8.9 Averaged results of Study 4

	Average percentage of agreement with both sentences (≥4)
Target pairs	23.5
Control pairs	0.5

Table 8.10 Gender and personality results of Study 4 (ns: nonsignificant)

Pair		A	B	D	E	G	H
Demographic variables	Gender	ns	ns	ns	ns	ns	ns
	Openness	ns	ns	ns	ns	ns	ns
	Conscientiousness	ns	ns	ns	ns	46/0	ns
	Extraversion	ns	ns	ns	ns	ns	ns
	Agreeableness	ns	ns	ns	ns	ns	ns
	Neuroticism	ns	ns	ns	ns	ns	ns

5.4 Discussion

In Study 4 as in Studies 1 and 2, which used different stimuli, participants are much more likely (nearly 50 times) to agree with the two sentences of the target pairs than with the two sentences of the control pairs. No clear variation across domains emerges, but, tentatively, pairs about biological kinds seem less likely to elicit agreement to both sentences than other pairs (artifacts and human activities). The proportion of participants willing to assent to both sentences of the target pairs remains small (roughly a quarter of our sample), although it is larger than in Study 2.

Study 4 replicates the main finding of Study 3 with new stimuli and a larger sample size: There is no evidence of systematic personality-related or gender-related variation in people's willingness to embrace apparently contradictory sentences.

6 Conclusion

There are three main lessons to be drawn from the present extension of the line of research first developed in Machery and Seppälä (2010). First, the phenomenon reported in Machery and Seppälä is robust: A substantial proportion of competent speakers are willing to assent to apparently contradictory sentences, suggesting that the relevant words ("vegetable," "butter," etc.) can express different concepts. This phenomenon was replicated in this chapter with the original stimuli and with new stimuli.

Second, for most stimuli, most people are reluctant to assent to apparently contradictory sentences. This finding can be explained in light of the results of Study 2 of this chapter. As we saw in Study 2, when the hedge "in a sense" is not used, only a small proportion of participants are willing to assent to apparently contradictory sentences. This suggests that, while several distinct coreferential concepts can be expressed by a particular word (e.g., a prototype and a causal theory of fish can be expressed by "fish"), one of these concepts is preferentially expressed by this word, which is by default understood as expressing it. A hedge like "in a sense" is needed to prime people to express the other coreferential concepts. This proposal can explain why most people

do not assent to apparently contradictory sentences: Many people stick with the concept preferentially expressed by a word.

Third, no evidence was found that gender or any of the five fundamental personality traits (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism) influences whether one is disposed to assent to two apparently contradictory sentences. This is particularly the case of conscientiousness, which a priori was a plausible factor influencing this disposition. It remains unclear what factors lead people to assent to two apparently contradictory sentences.

So, how does the research reported in this chapter bear on the polysemy hypothesis and on the heterogeneity hypothesis? First, it supports a revised version of the polysemy hypothesis. The original polysemy hypothesis proposes that many words are polysemous because they express several distinct coreferential concepts. This is inconsistent with the findings of Study 2, which suggests that words preferentially express a single concept. The revised polysemy hypothesis proposes that many words *can* express several distinct concepts. Hedges such as "in a sense" allow speakers to replace the preferred interpretation of a word with another (nonmetaphorical) one. Speakers are more or less likely to follow through with this substitution depending on yet unknown factors. This revised polysemy hypothesis supports the heterogeneity hypothesis: That words can express several distinct coreferential concepts supposes the existence of several distinct coreferential concepts.

Notes

1 For discussion, see Machery 2006, 2011; Piccinini and Scott 2006; Weiskopf 2009; Hill 2010; Malt 2010; Prinz 2010; Lombrozo 2011; Piccinini 2011; Poirier and Beaulac 2011; as well as the commentaries on Machery 2010.

2 Note that the heterogeneity hypothesis is not entailed by the polysemy hypothesis since different occurrences of a given word could express different concepts because different concepts are constructed on the fly, depending on the context in which this word occurs. This contextualist view of concepts, defended, for example, by Barsalou and Prinz, is inconsistent with the heterogeneity hypothesis (for discussion, see Machery forthcoming).

3 In what follows, for the sake of simplicity, I will only focus on classes.

4 Their design is described in more detail in Study 1 below (Section 2.1).

5 Thus, the first sentence of each pair was affirmative, while the second was negative.

References

Gosling, S. D., Rentfrow, P. J., and Swann, W. B., Jr. (2003), "A very brief measure of the big five personality domains." *Journal of Research in Personality*, 37, 504–28.

Hill, C. (2010), "I love Machery's book, but love concepts more." *Philosophical Studies*, 149, 411–21.

Lombrozo, T. (2011), "The campaign for concepts." *Dialogue*, 50, 165–77.

Machery, E. (2005), "Concepts are not a natural kind." *Philosophy of Science*, 72, 444–67.

—(2006), "How to split concepts. Reply to Piccinini and Scott." *Philosophy of Science*, 73, 410–18.

—(2009), *Doing without Concepts*. New York: Oxford University Press.

—(2010), "Précis of *Doing without Concepts*." *Behavioral and Brain Sciences*, 33, 195–244.

—(2011), "Replies to Lombrozo, Piccinini, and Poirier and Beaulac." *Dialogue*, 50, 195–212.

—(forthcoming), "By default: Concepts are accessed in a context-independent manner," in E. Margolis and S. Laurence (eds), *Concepts II: A Reader*. Cambridge, MA: MIT Press.

Machery, E. and Seppälä, S. (2010), "Against hybrid theories of concepts." *Anthropology & Philosophy*, 10, 97–125.

Malt, B. (2010), "Why we should do without concepts." *Mind & Language*, 25, 622–33.

Piccinini, G. (2011), "Two kinds of concept: Implicit and explicit." *Dialogue*, 50, 179–93.

Piccinini, G. and Scott, S. (2006), "Splitting concepts." *Philosophy of Science*, 73, 390–409.

Poirier, P. and Beaulac, G. (2011), "Le véritable retour des définitions." *Dialogue*, 50, 153–64.

Prinz, J. (2010), "Can concept empiricism forestall eliminativism?" *Mind & Language*, 25, 612–21.

Sennet, A. (2011), "Ambiguity," in E. Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Summer 2011 Edition). http://plato.stanford.edu/archives/sum2011/entries/ambiguity/.

Weiskopf, D. (2009), "The plurality of concepts." *Synthese*, 169, 145–73.

Zwicky, A. M. and Sadock, J. M. (1975), "Ambiguity tests and how to fail them." *Syntax and Semantics*, 4, 1–36.

Index

Adam, D. 118
Agency and Experience
as sources of moral standing 115–18, 119n. 3
subclinical populations, dissociation in 114–15
agency model of mental state ascriptions 5
Aggression Questionnaire 145
alarm clock 103
"alarm-like" emotional responses 144
analytic reasoning 133
connection between pro-social 163
analytic thinking 126
animalist conception, of self 189
animalistic dehumanizing narrative 138
anterior medial cingulate cortex (aMCC) 111
anti-intuitionist thesis 26, 28, 30, 34
Arico, A 5, 55–6, 58
autism 28, 109, 113, 129
Autism Spectrum-Quotient Scale 115
autistic spectrum disorders (ASD) 109–10
Aydede, M. 76, 81, 90, 92–4, 96nn. 1–3

Bartels, D. 184–5, 199n. 4
Bealer, G. 27–8, 31–3
Bering, J. M. 50–1, 69n. 9
biological conception, of self 189, 194
biological continuity 199n. 2
compensation 190
social treatment 189
biological embodiment 46
unified 47–9
Bjorklund, D. 51
Block, N. 2–4, 12–14, 45–6, 48, 68, 78
Bonjour, L. 25
Buckwalter, W. 68

calibration arguments of intuitions 18–22
Cappelen, H. 37
causal theories 203–7, 209–10, 214, 219
China brain intuition 12–15, 45, 48
Choe, S. Y. 141
cognitive empathy 108–10
Cognitive Reflection Test 146
commonsense conceptions 193
concepts, class of 203
conceptual analysis 33–4
conceptual monism 7
connectedness 199n. 3
beliefs about 199n. 4
effect on anxiety 186–7
manipulation, and judgments about punishment 185–6
consciousness 1, 3–4, 14, 17–19, 36, 47, 55, 68
contradiction test 206–7
Côté, S. 141
cue-based paradigms 111
cultural variation arguments, of intuitions 15–17
Cummins, R. 18–19

DANVA facial emotion recognition task 145
default mode network 130
De Greck, M. 110
dehumanization 160–1
dehumanized out-groups 160
Dennett, D. 6, 48, 101, 104, 108, 129
deontological judgments 127–8
deontological morality 115
deontological reasoning 148–50
scenarios to conform with, study of 148–50
vignettes 167–71
determinate concept possession 32
Deutsch, M. 36–7
Devitt, M. 22–4
"differing background beliefs" hypothesis 17, 39n. 5
disembodied entities 50–1

disembodying ascriptions, experiments investigating
ascribing emotional states to ghosts and spirits 62–7
ascriptions of emotion to ghosts 52–5
with eternally disembodied spirits 58–62
lacking intentional objects with biasing contextual information 55–8
Doing without Concepts 203
dual-process cognitive system 5
dual-process theory 126–8, 133
Dunfield, K. A. 107
dyadic completion 114

East Asian knowledge 17
embodiment hypothesis 4, 46–8, 51
of folk psychological judgments 47
emotion on moral judgment, effects on 131–2, 139–41
Emotion Regulation Questionnaire 145
Emotional and Social Competency Inventory 145
emotional contagion 109
Emotional Empathic Tendency Scale 141
emotional empathy 108, 110–11
empathetic reasoning 133
empathy, psychological constructs of 141–2
empirical case, for pluralism *see also* pluralism, about personal identity
manipulation studies 184–7
self, different senses of 187–8
survey studies 183–4
empirical evidence 181
episodic memory 187–8
epistemological intuitions 26–8
euthanasia judgments 143
active euthanasia, scenarios involving 156–8
passive euthanasia, scenarios involving 153–6
resistance to 161, 172n. 8
exemplars 203–5
experiential mental states 130
experimental philosophy 1–2, 11–12, 22, 29, 38–9
external corroboration 21
extrapolative calibration 20

Fiala, B. 5
Five Facet Mindfulness Questionnaire 145
folk pluralism *see* pluralism, about personal identity
folk psychology 22, 46
Friedman, J. 6
From Metaphysics to Ethics 33
frontotemporal lobar degeneration 109
functional information, influence on mental states 49–50, 68
functionalist theory of mind 13

Gaita, R. 160–1
gender, and individual variation study 215–18
Gettier intuitions 15–16
Gleichgerrcht, E. 141, 143, 149–50, 159
Goldman, A. 20, 34–5
Gray, H. M. 2, 6, 50, 115
Gray, K. 47, 69n. 7, 114–15
Greene, J. D. 6, 127–8, 133–4, 136–42, 144, 147–8, 156, 159–60, 163–4, 170, 172nn. 3–4

hallucinations of pain 78–9, 83–95 *see also* pain
Aydede's challenge 92–5
between participants 85–8
vs pain illusions 88–90
understanding 90–2
within participants 83–5
Hamlin, J. K. 106
Harman, G. 26
Heider, F. 106
heterogeneity hypothesis 7, 203–5, 215, 220n. 2
domain-generality of 210–11
individual variation study 215–17
limits of study 208
Machery and Seppälä's exploratory study 207–8
new stimuli, extending study with 217–19
replication of Machery and Seppälä's study (*see* Machery and Seppälä's study, replication of)
in a sense, role in 213–15
target and control sentences 207

Hill, C. 76, 81, 89, 96nn. 1–2
hopefulness 30
hopelessness 30
Howard, C. 7
Huebner, B. 14, 112
humanity, principle of 160

identity-related practical concerns 181
impaired cognitive empathy 109
implied ontology 34
Imposing Memory Task 145
in a sense, role of 213 *see also* heterogeneity hypothesis
  findings interpretation, supports heterogeneity hypothesis 215
  polysemy hypothesis revision 214, 220
indispensability arguments of intutions 25–9
individual differences, moral reasoning and 139–41, 144–8
informational/phenomenal objection 5
intentional stance 6, 101–8, 129 *see also* phenomenal stance; phenomenological stance
  adoption of 110
  automatic nature of attribution of 106
  bottom-up elicitation of an 106
  dissociation of phenomenological stance and 108–9, 112
  early emergence of 106–7
  example of a chess-playing computer 104–5
  to predict behavior 105–6
  to understand complex behaviors 104–5
intentional states 47
Interpersonal Reactivity Index (IRI) 145–6
intuition trump theory 24
intuitions 2 *see also* mental states; moral reasoning
  about group agents 46–7
  about privacy and subjectivity 79–82
  analogy between observation and 15–17
  arguments against use of 36–7
  basic vs rich 23
  calibration arguments 18–22
  constitutivity arguments 31–6
  "contaminated" by theory 23
  cultural variation arguments 15–17
  Cummins's ultimate conclusion 18–20
  degree of evidential weight 13
  degree of psychological heterogeneity in 28
  degree of theoretical illumination of 22
  epistemological similarity between perception and 29
  explanationist criterion and 27
  false presupposition arguments 36
  Goldmans's claims 20
  indispensability arguments 25–9
  influence of background theory 24
  as "introspectively opaque" 22
  parity arguments 29–31, 36
  and personal identity 7
  in philosophy of mind 12–15
  presupposition arguments 36–9
  psychological mechanisms underlying 22
  restriction arguments 22–5
  "traditional" intuition-based argumentation 9
  use in philosophy 24
  variation arguments 15–18
  Western subjects vs East Asian subjects 15–16

Jack, A. I. 6–7, 108
Jackson, F. 33–5, 41n. 12, 193–4
justification, explanationist criterion of 26
justified true belief (JTB) model of knowledge 24

Kant, I. 19, 115–16, 132, 138, 160, 163–4
Kasparov, G. 104–5
Knobe, J. 2–5, 14, 35, 46–7, 69n. 3, 112, 119nn. 2, 4
Koenigs, M. 146, 149, 151
Kornblith, H. 22, 24
Kripke, S. 37, 78
Kuhlmeier, V. A. 107
Kumar, V. 7
Kvaran, T. 7

Lewis, D. 77, 96n. 3

Machery and Seppälä's study, replication of 207–8
  biographical questionnaire, includes personality inventory 211
  lessons drawn from extention of 219–20
  scoring and data analysis 211–12
  subjects and methodology 209–11
  target and control sentences 211
  variation across pairs similar to 212
Machery, E. 3–8, 11, 36, 103, 115, 119n. 2, 203–4, 207–9, 211–14, 219
manipulation studies, affect on judgments
  anxiety, and 184–5
  connectedness of self 184–5
  economic decisions 186
  judgments about punishment 185
  Williams's key thought experiment 186
mechanistic dehumanizing narrative 138
mental states 2–3 *see also* intuitions; moral reasoning
  attribution using a dimensional approach 113
  attributions to disembodied entities 50–1
  dimensions, agency and experience 102–7, 113–15
  functional information, influence of 49–50
  moral rights and responsibilities 114
  role of embodiment 46
Meyers, C. 6
Microsoft Corporation 46–7
Min, K-H. 141
mind perception, dimensions of 113–15
monistic views 181, 198
monolithic intuition 28
moral agency, judgments of 6
moral cognition 126, 128, 131, 134, 138
moral intuition 26–7, 136
moral judgment 127–8
  dual-process theory of 128
moral patiency
  judgments of 6
  phenomenal experience perception and 172n. 6
moral personhood 107–8
moral reasoning 27
  analytic brain areas involved in 136
  behavioral evidence 141–2
  deontological principle scenarios and 148–50, 167–71
  differences between the models and neuroimaging evidence 132–9
  direct personal-harm scenarios and 144–8, 165–7
  emotions and 131–2, 139–41
  Greene's view of 172n. 4
  language of 161–2
  medial moral regions involved in 135–7
  midline moral regions involved in 137–8
  neuroimaging of individual differences 139–41
  relationship between compassion and resistance to utilitarian reasoning 150–3
multigons 31–2

"nation of China" thought experiment 2, 42, 68
"negative" experimental philosophy 15
neuroimaging of morality 128
Nichols, S. 5, 35, 67n. 9
nociception 77
nominal concept possession 32
nonintuitive justification 27
nonsocial reasoning 128

Occam's razor 35
OCEAN model 215
Olson, E. 195–6
opposing domains theory 133
ownership 191–2

pain
  appearance-reality distinction of 82–3
  awareness of 76–9
  common-sense/folk conception of 77, 80–1, 96n. 3
  expressions of 81, 94–5
  hallucinations 78–9, 83–95
  introspective accounts of 77–9
  matrix 111–12
  ordinary conception of 81
  paradox of 76–9

perceptual account of 76–7
privacy and subjectivity of 79–82
processing of pain stimuli 77
parasympathetic nervous system 140
Parfit, D. 184, 199n. 3
parity arguments of intuitions 29–31, 36
passion, fundamental tension between reason and 125–6
perceptions 21
epistemological similarity between intuition and 29
parity between intuition and 30
personal identity *see also* pluralism, about personal identity
biological approach to 182, 193–4, 198
different views of 182–3
and experimental philosophy 181
monism role 181
pluralistic approach to 195–6
pluralist philosophical view about 181
and psychological connectedness 199n. 3
psychological traits view 182–3, 187, 189, 194, 198
subject matter of 193–4
personality traits, and individual variation study 215–18
Phelan, M. 4, 63, 68, 72n. 25
phenomenal consciousness 3, 47, 69n. 2
phenomenal stance 6, 161, 163
differences between the models and neuroimaging evidence 132–9
phenomenal state ascriptions to entities 52–63, 66, 68
ascribing emotional states to ghosts and spirits 62–7
lacking intentional objects with biasing contextual information 55–8
with eternally disembodied spirits 58–62
phenomenally conscious mental states 3
phenomenological stance 108 *see also* intentional stance; phenomenal stance
adoption of 110–13
dissociation of intentional stance and 108–9, 112
Phillips, D. 5
philosophical intuition 30
philosophical thought experiments 181
physical stance 6, 103, 129, 163
pluralism, about personal identity
biological approach to identity 182, 198
coherence of 195–7
empirical case of (*see* empirical case, for pluralism)
identity vs ownership 190–3
ownership concept 190
philosophical, monism price 193–5
and practical concerns 188–9
psychological traits view in identification 182, 198
pull of 197–8
Shoemaker's argument, practical concerns 189–3
thought experiments in 181–2
too-many-thinkers problem 195–6
polysemy hypothesis 7, 204–5, 210, 220
contradiction test 206
importance of hedges and 214
tests for 206–7
posterior anterior cingulate cortex (pACC) 111
presupposition arguments of intuitions 36–9
Prinz, J. 2–5, 14, 46–7, 69nn. 3–4, 112, 119n. 2
pro-social emotions 140–2
influence on moral reasoning 143–4, 148–50
prototypes 203–10, 214, 219
prototypical sentences 210–11, 217
psychological concepts 34–5
psychological connectedness 184, 199n. 3
psychological continuity 189–90, 199n. 3
psychological traits 182–3, 187–9, 196, 198, 199n. 2
psychopathy 28, 109–10, 129, 141
Pust, J. 26–7
Putnam, H. 78

Reading the Mind in the Eyes Task 145
reason, fundamental tension between passion and 125–6
reliable belief-forming mechanism 16
Reuter, K. 5, 79, 82–3, 95
rich intuitions 23–4

Rips, L. 183
Robbins, P. 6, 108

Schizotypy Personality Questionnaire 115
Schwitzgebel, E. 80
scientific reasoning 130
"sees red" 4–5
self conceptions 186
Self-Report Psychopathy Scale (SRP)-III 145–6
semantic memory 187
Seppälä, S. 7, 203–4, 207–9, 211–14, 219
Shoemaker, D. 189–92, 197, 200
  identity vs ownership case 190–3
  views on pluralism of practical concerns 188–9
Sider, T. 182
Silver, N. 105
Simmel, M. 106
simulation constraint hypothesis 69n. 9
social cognition 131, 134, 137
social emotions 131, 144–5, 172n. 6
social interaction, mode of 131
social reasoning 128, 138–40
Social Stories Questionnaire 145
Sosa, E. 16–17, 29–30, 39n. 5
Stich, S. 17
subclinical populations, dissociation of Agency and Experience in 114–15
suffering index 114
supernatural language of morality 161–2
Swain, S. 16, 39n. 5
sympathetic nervous system 140–1
Sytsma, J. 3–6, 79–81, 83, 95, 103, 115, 119n. 2

task-positive network 130, 137
Ten Item Personality Measure 211
theoretical sentences 210
theory-contamination 23
theory of mind 1, 101, 109–10, 140
Theriault, J. 6
Tierney, H. 7
trait view, of personal identity 182
"Troubles with Functionalism" 12–13
Tye, M. 90, 96nn. 2–3

Urminsky, O. 184–5, 199n. 4
utilitarianism 115, 139
  Greene's conception 163–4
  Kant's conception 163–4
  link between emotionality and 142
  normative significance 163–4
  pro-social behavior and 164
utilitarian moral judgments 133
utilitarian reasoning 131
  negative relationship between compassion and 153–6
  relationship between compassion and resistance to 150–3
utilitarian thinking 131, 142

variation argument of intuitions 15–18
ventromedial prefrontal cortex (vMPFC) 139–40

Waytz, A. 105–6, 119n.1
Weatherson, B. 22, 24–5, 39n. 3
Wegner, D. M. 69n. 7, 114
Weinberg, J. 11, 15–17, 20–2, 30–1, 39n. 1, 40n. 9
Werning, M. 83, 96n.2
Wesley, B. 4
What are We? 195
Williams, B. 181–3, 186–8, 198
Williamson, T. 30, 36
Williams-style case 186–7

Young, L. 6, 141, 143, 149–50, 159